The Dissertation Committee for Jason Jo
certifies that this is the approved version of the following dissertation:

Structured Low Complexity Data Mining

Committee:

Rachel Ward, Supervisor

Peter Mueller

Ronny Hadani

Kui Ren

James Scott

Structured Low Complexity Data Mining

by

Jason Jo, B.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2015

Acknowledgments

First I would like to thank my family for loving and supporting me unconditionally throughout my entire life. I would like to thank my adviser Rachel Ward for her guidance and support. I would also like to thank Peter Müller, who taught some amazing courses and has sparked my fascination with Bayesian Statistics. I thank my dissertation committee for their service and their helpful feedback. And finally, I am grateful for all the wonderful people in the math department, from the grad students to the extremely helpful and the fantastic staff to all the great professors.

Structured Low Complexity Data Mining

Publication No. _____

Jason Jo, Ph.D.
The University of Texas at Austin, 2015

Supervisor: Rachel Ward

Due to the rapidly increasing dimensionality of modern datasets many classical approximation algorithms have run into severe computational bottlenecks. This has often been referred to as the "curse of dimensionality." To combat this, low complexity priors have been used as they enable us to design efficient approximation algorithms which are capable of scaling up to these modern datasets. Typically the reduction in computational complexity comes at the expense of accuracy. However, the tradeoffs have been relatively advantageous to the computational scientist. This is typically referred to as the "blessings of dimensionality."

Solving large underdetermined systems of linear equations has benefited greatly from the sparsity low complexity prior. A priori, solving a large underdetermined system of linear equations is severely ill-posed. However, using a relatively generic class of sampling matrices, assuming a sparsity prior can yield a well-posed linear system of equations. In particular, various greedy

iterative approximation algorithms have been developed which can recover and accurately approximate the k-most significant atoms in our signal. For many engineering applications, the distribution of the top k atoms is not arbitrary and itself has some further structure. In the first half of the thesis we will be concerned with incorporating some a priori designed weights to allow for *structured* sparse approximation. We provide performance guarantees and numerically demonstrate how the appropriate use of weights can yield a simultaneous reduction in sample complexity and an improvement in approximation accuracy.

In the second half of the thesis we will consider the collaborative filtering problem, specifically the task of matrix completion. The matrix completion problem is likewise severely ill-posed but with a low rank prior, the matrix completion problem with high probability admits a unique and robust solution via a cadre of convex optimization solvers. The drawback here is that the solvers enjoy strong theoretical guarantees only in the uniform sampling regime. Building upon recent work on non-uniform matrix completion, we propose a completely expert-free empirical procedure to design optimization parameters in the form of positive weights which allow for the recovery of arbitrarily sampled low rank matrices. We provide theoretical guarantees for these empirically learned weights and present numerical simulations which again show that encoding prior knowledge in the form of weights for optimization problems can again yield a simultaneous reduction in sample complexity and an improvement in approximation accuracy.

Table of Contents

Acknowledgments iv

Abstract v

List of Figures ix

Chapter 1. Introduction 1

Chapter 2. Structured Sparse Solutions to Underdetermined Systems of Linear Equations 4

2.1 Overview 4
2.2 Sparsity as a Low Complexity Prior 6
2.3 RIP Matrices and Convex Relaxation 11
2.4 Iterative Hard Thresholding 15
2.5 Weighted Sparsity 18
2.6 Iterative Hard Weighted Thresholding: In Theory 21
2.6.1 Extension to the Weighted Case 21
2.6.2 Performance Guarantees 23
2.6.2.1 Performance Guarantees: Convergence to a Neighborhood 23
2.6.2.2 Performance Guarantees: Contraction 31
2.7 Iterative Hard Weighted Thresholding: In Practice 35
2.7.1 Choosing the weights 35
2.7.2 Approximate Projection 36
2.7.3 Experiments 37
2.8 Conclusion and Future Directions 43

Chapter 3. Collaborative Filtering: Weighted Matrix Completion **46**
3.1 Overview . . . 46
3.2 Nuclear Norm Minimization and Uniform Sampling . . . 50
3.3 Non-uniform Matrix Completion . . . 55
3.4 Main Results . . . 58
3.5 Empirical Estimation . . . 63
3.5.1 Proof Lemma 3.4.1 . . . 65
3.5.2 Proof of Lemma 3.4.2 . . . 67
3.6 Matrix Completion Guarantees . . . 69
3.6.1 Proof of Theorem 3.4.3 . . . 69
3.6.2 Weighted Nuclear Norm and Relaxation of Sufficient Recovery Conditions . . . 70
3.6.2.1 Proof of Theorem 3.4.4 . . . 70
3.7 Numerical Experiments . . . 72
3.8 Conclusion . . . 74

Chapter 4. Alternate Weighted Matrix Completion Analysis **76**
4.1 Introduction . . . 76
4.2 Main Results . . . 77
4.3 Empirical Estimation . . . 82
4.3.1 Proof of Lemma 4.2.1 . . . 83
4.3.2 Proof of Lemma 4.2.3 . . . 89
4.4 Matrix Completion Guarantees . . . 92
4.4.1 Proof of Theorem 4.2.2 . . . 92
4.4.2 Weighted Nuclear Norm and Relaxation of Sufficient Recovery Conditions . . . 93
4.4.2.1 Proof of Theorem 4.2.4 . . . 94

Chapter 5. Conclusion **96**

Bibliography **97**

List of Figures

2.1 Exact Recovery of Randomly Generated Variable s-sparse Power Law Signals using $m = 128$ measurements. Results are averaged over 200 trials. Best viewed in color. 39

2.2 Exact Recovery of a fixed sparse $s = 25$ power law distributed signal using a variable number of measurements. Results are averaged over 200 trials. Best viewed in color. 40

2.3 The log normalized error averaged over 200 trials of noisy s-sparse approximation of dense Power Law Signals using $m = 128$ measurements. Best viewed in color. 41

2.4 The log standard deviation of 200 trials of noisy s-sparse approximation of dense Power Law Signals using $m = 128$ measurements. Best viewed in color. 42

2.5 The log normalized error averaged over 200 trials of noisy s-sparse approximation of dense Power Law Signals using a variable number of measurements. Best viewed in color. 43

2.6 The log standard deviation of 200 trials of noisy s-sparse approximation of dense Power Law Signals using a variable number of measurements. Best viewed in color. 44

3.1 Probability of Exact Recovery when the rank is equal to 5. . . 73

3.2 Probability of Exact Recovery when the rank is equal to 10. . 73

3.3 Probability of Exact Recovery when the rank is equal to 15. . 74

3.4 Probability of Exact Recovery when the rank is equal to 20. . 74

3.5 Probability of Exact Recovery when the rank is equal to 25. . 74

3.6 Power Law Sampling with replacement rate vs. Percentage of Unique Samples Revealed. 74

Chapter 1

Introduction

As the dimensionality of modern data has exploded, many classical approximation methods have become intractable. For this thesis, we will consider two such tasks:

- *Solutions to underdetermined systems of linear equations*:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon} \text{ where } \boldsymbol{y}, \boldsymbol{\epsilon} \in \mathbb{C}^m, \boldsymbol{A} \in \mathbb{C}^{m \times n}, \boldsymbol{x} \in \mathbb{C}^n, \tag{1.1}$$

  where we consider the high dimensional regime in which $m < n$.

- *Matrix completion*: For a given data matrix $\boldsymbol{M} \in \mathbb{R}^{m \times n}$, let Ω denote a set of observed entries of $\boldsymbol{M}$ where $\Omega \sim \boldsymbol{p}$ and $\boldsymbol{p}$ is a probability mass function on $[m] \times [n]$. Typically the cardinality of Ω is a small fraction of the $m \cdot n$ total entries. The matrix completion task is the following problem: given the known entries Ω, infer the remaining entries $[m] \times [n] \setminus \Omega$ of $\boldsymbol{M}$.

Clearly the above two problems are severely ill-posed in their respective settings. Typically a low complexity prior hypothesis is assumed to make such problems tractable. Compressed Sensing [1] is the study of solutions to

(1.1) which are *sparse*. The first generation of compressed sensing algorithms operated under the implicit assumption that all sparsity patterns are equally likely. However there are industrial problems in which certain sparsity patterns occur with a higher frequency than other sparsity patterns. This phenomenon has been observed in a more generalized form of the frequent occurrence of power law distributions in empirical data [2]. Similar phenomenon exists in datasets which are stored as matrices, for example in e-commerce with user-product ratings. Many times a small percentage of users are responsible for a large percentage of ratings and opinions, so-called influential "power-users."

In the first half of this thesis we will be interested in exploiting prior knowledge of a signal's energy distribution encoded in the form of optimization weights to simultaneously reduce sample complexity and improve approximation accuracy. *While weighted ℓ_1 minimization methods have been developed which enjoy structured sparse approximation guarantees, these methods all suffer from the fact that they scale poorly to the high dimensional setting.* In this case, typically one does a tradeoff of accuracy for computational efficiency and appeals to iterative greedy approximation methods. To this end, *we will modify a greedy efficient iterative approximation algorithm called Iterative Hard Thresholding (IHT) to incorporate weights and provide theoretical performance guarantees for both the exact (i.e. $\boldsymbol{\epsilon} = 0$ in (1.1)) and noisy cases.*

For the matrix completion problem, typically one assumes a low complexity prior on the data matrix $\boldsymbol{M}$ either being low rank or being well approximated by a low rank matrix. A matrix having low rank is again a form

of sparsity; indeed the rank of a matrix is equivalent to the sparsity of the vector of a matrix's singular values. Miraculously, assuming a low rank prior allows for the matrix completion task to be relaxed to a convex optimization problems which admits a cadre of tractable solution methods. The drawback is that these matrix completion solvers only have theoretical guarantees in the uniform sampling regime, i.e. Ω is a uniform sample of $[m] \times [n]$. To this end, recent work by [3] established that one can design optimization weights which depend on the sampling distribution $\boldsymbol{p}$ which will allow for the exact recovery of any arbitrarily sampled low rank matrix. *In general, the sampling distribution $\boldsymbol{p}$ is not known to us and one may consider the aforementioned weights to be of the same class of weights considered in the first half of this thesis: expertly designed.* To this end, in the second half of this thesis, *we will provide theoretical guarantees for a set of empirically learned weights which assumes zero expert or prior knowledge but allows for exact recovery of arbitrarily sampled low rank matrices.*

Therefore, in this thesis we aim to explore the two complementary sides of using prior knowledge to improve performance of high dimensional approximation algorithms: (1) expert knowledge is encoded in the form of a priori weights and (2) no expert knowledge is available and instead we use statistically principled estimators to learn sufficient weights. In both cases, we will present numerical simulations which provide evidence that the appropriate use of weights for high dimensional approximation tasks can yield a simultaneous reduction in sample complexity and improvement in approximation accuracy.

Chapter 2

Structured Sparse Solutions to Underdetermined Systems of Linear Equations

2.1 Overview

Compressed sensing algorithms attempt to solve underdetermined linear systems of equations by seeking structured solutions, namely that the underlying signal is either sparse or well approximated by a sparse signal [1]. However, in practice much more knowledge about a signal's support set is known beyond that of sparsity or compressibility. Empirically it has been shown that the spectral power of natural images decays with frequency f according to a power-law $1/f^p$ for $p \approx 2$ [4, 5]. Likewise, the frequency of earthquakes corresponding to their magnitudes as measured by Moment magnitude scale empirically also exhibits a power law decay [6]. For these types of highly structured signals, certain atoms in the dictionary are more prevalent in the support set of a signal than other atoms. The traditional notion of sparsity treats all atoms uniformly and thus is not ideally suited to utilize this rich prior knowledge.

To this end one can consider using weighted ℓ_1 minimization to obtain structured sparse solutions. Weighted ℓ_1 minimization can leverage prior

knowledge of a signal's support to undersample the signal, and avoid overfitting the data [7–10]. However, *the main drawback that weighted ℓ_1 minimization shares with ℓ_1 minimization is that solution methods scale poorly* [11].

While many computationally efficient approximation algorithms have been developed for computing a best s-sparse approximation [1] no such method has been developed for the weighted case. In this chapter, we make the following contributions:

1. Using a generalized notion of weighted sparsity and a corresponding notion of Restricted Isometry Property on weighted sparse signals developed in [7], *we pose a weighted analogue of the best s-sparse approximation problem.*

2. **An extension of the Iterative Hard Thresholding (IHT) algorithm [12] is presented to solve the weighted sparse approximation problem.** We emphasize how the same template used to derive performance guarantees for all the greedy compressed sensing algorithms carries over naturally. Indeed, performance guarantees are derived and much of the theoretical principles remain intact. However, not all theoretical results extend and the barrier seems to be the nature of weighted thresholding. We explore this extension barrier and present a detailed analysis of which theoretical guarantees do not extend and how the barrier is responsible for this obstruction. Under an additional hypothesis, the extension barrier is rendered moot and we present some specialized

theoretical guarantees.

3. While both IHT and the IHT extension compute a projection onto a non-convex space, the projection that IHT requires can actually be efficiently computed while the projection that our IHT extension requires does not seem to have an efficient solution. To this end, we consider a tractable surrogate to approximate this non-convex projection and we present its empirical performance on power law distributed signals.

2.2 Sparsity as a Low Complexity Prior

In this section we will develop the sparsity prior in the context of recovering an n-dimensional signal $\boldsymbol{x}^*$ from a (possibly noisy) set of m linear measurements $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{\epsilon}$ where $m < n$. Consider first the noiseless case when $\boldsymbol{\epsilon} = 0$. We may pose the Ordinary Least Squares (OLS) problem:

$$\min_{\boldsymbol{x}\in\mathbb{C}^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2. \tag{2.1}$$

Note that (2.1) allows us to also consider the noisy case. As $m < n, \boldsymbol{A}$ has a non-trivial null-space and there there are infinitely many solutions to (2.1). The objective function of (2.1) $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$ is differentiable and has a symmetric positive-semidefinite Hessian matrix $\nabla^2 f(\boldsymbol{x}) = \boldsymbol{A}^*\boldsymbol{A} \succeq 0$. Note that $\nabla^2 f(\boldsymbol{x})$ cannot be strictly positive definite as $\boldsymbol{A}$ has a non-trivial null-space. To this end, (2.1) is a convex problem with infinitely many global optimum. When $\boldsymbol{A}$ is full rank, one may obtain an explicit solution to (2.1)

given by:

$$\boldsymbol{x}^+ := \boldsymbol{A}^+\boldsymbol{A}^*\boldsymbol{y}, \tag{2.2}$$

where $\boldsymbol{A}^+ := (\boldsymbol{A}^*\boldsymbol{A})^{-1}$ is the *Moore-Penrose Pseudo-inverse.* Using the compact Singular Value Decomposition of $\boldsymbol{A} = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^*$ where $\hat{\boldsymbol{U}} \in \mathbb{C}^{m\times m}, \hat{\boldsymbol{\Sigma}} \in \mathbb{C}^{m\times m}, \hat{\boldsymbol{V}}^* \in \mathbb{C}^{m\times n}$ we have that:

$$\boldsymbol{x}^+ = \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^*\boldsymbol{x}^*, \tag{2.3}$$

in other words that the solution $\boldsymbol{x}^+$ is the projection of the true signal $\boldsymbol{x}^*$ onto the first m right singular vectors of the sensing matrix $\boldsymbol{A}$. To this end we may conclude that $\boldsymbol{x}^+$ is typically a poor approximation to $\boldsymbol{x}^*$. One may take a mirror view of (2.1) as a linear regression problem, as opposed to a signal approximation problem. In this sense, the approximation $\boldsymbol{x}^+$ represents linear regression coefficients and each row of $\boldsymbol{A}$ is merely a data input value. In the context of regression, we have the bias-variance tradeoff. When we have that the error term $\boldsymbol{\epsilon}$ satisfies homoscedasticity and its components are uncorrelated, then by the Gauss-Markov Theorem [13] the bias is zero, however our variance is likely to be very high as noted before that $\boldsymbol{x}^+$ will in all likelihood be a poor approximation of $\boldsymbol{x}^*$ and will in all likelihood predict poorly on new data points.

One modification we can make to (2.1) is to add a *regularization term* so that the problem becomes well-posed, i.e. there exists a numerically stable and unique solution. The first regularizer we consider is the *Tikhonov regularizer*

which imposes an ℓ_2 penalty on the signal approximation:

$$\min_{\boldsymbol{x}\in\mathbb{C}^n} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{x}\|_2^2, \lambda > 0. \tag{2.4}$$

Note that this unconstrained optimization problem is equivalent to the following constrained optimization problem for some $t > 0$ which depends on λ:

$$\min_{\boldsymbol{x}\in\mathbb{C}^n} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 \text{ subject to } \|\boldsymbol{x}\|_2 \leq t. \tag{2.5}$$

The optimization problem (2.4) is referred to as *Ridge Regression*. Observe that the objective function for (2.4) $f_\lambda(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{x}\|_2^2$ has Hessian $\nabla^2 f_\lambda(\boldsymbol{x}) = (\boldsymbol{A}^*\boldsymbol{A} + \lambda\boldsymbol{I})$ which is a strictly positive definite matrix. Therefore the ridge regression problem is strongly convex and thus admits a unique solution, which has closed form:

$$\boldsymbol{x}_\lambda = (\boldsymbol{A}^*\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}^*\boldsymbol{y}. \tag{2.6}$$

Using the compact form of the SVD for $\boldsymbol{A}$ we have that:

$$\boldsymbol{x}_\lambda = \sum_{i=1}^{m} v_i \frac{s_i}{s_i^2 + \lambda}\langle u_i^*, \boldsymbol{y}\rangle, \tag{2.7}$$

where $\{s_i\}_{i=1}^m, \{v_i\}_{i=1}^n$ and $\{u_i^*\}_{i=1}^m$ denote the singular values, the columns of $\boldsymbol{V}$, and the rows of the matrix $\boldsymbol{U}$ SVD of $\boldsymbol{A}$, respectively.

Note that ridge regression shrinks the coefficients of the projection of $\boldsymbol{x}^*$ more along the smaller right singular directions of $\boldsymbol{A}$ and shrinks the coefficients of the aforementioned projection less along the larger right singular directions of $\boldsymbol{A}$ (corresponding to smaller/larger singular values).

Revisiting our mirror regression context, $\boldsymbol{x}_\lambda$ is now a biased estimator of $\boldsymbol{x}^*$, however the variance of our estimator depends on λ. Observe that when $\lambda = 0$ we recover the OLS estimate. For non-zero λ, observe that we will have non-zero shrinkage, with the hope that it will reduce the variance of our estimator $\boldsymbol{x}_\lambda$ by controlling how large our regression coefficients will grow.

While (2.4) is now a strongly convex optimization problem, the drawbacks of ridge regression include:

1. Still highly sensitive to the choice of sensing matrix $\boldsymbol{A}$.

2. Cannot zero out any components of a signal, i.e. it can only shrink, but for $\lambda \neq +\infty$, we are never eliminating variables and thus is incapable of performing *variable selection*. If we are in the high dimensional setting where $m \ll n$ this problem is further exacerbated and it becomes difficult to interpret $\boldsymbol{x}_\lambda$'s components.

To this end, we wish to consider a sparse regularizer to (2.1). We will be assuming the low complexity prior that $\boldsymbol{x}^*$ is either sparse or is itself well approximated by a sparse signal. Why assume a sparsity prior?

- Many natural signals are well approximated by sparse signals. Many standard data compression algorithms such as JPEG, MP3 and others exploit sparsity for example [1].

- Sparse signals have better interpretability.

- Sparse signal approximations have inherent low complexity and perhaps sparse approximation algorithms can be designed which scale gracefully in the high dimensional regime we are ultimately interested in. In fact, note that both the OLS solution and the ridge regression solution involve inverting a matrix or computing an SVD, both of which become computationally intensive for large dimensional data.

To this end, one may pose the following sparsity regularized OLS problem:

$$\min_{\boldsymbol{x}\in\mathbb{C}^n} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_0, \lambda > 0, \tag{2.8}$$

where $\|\cdot\|_0$ is referred to as the ℓ_0 pseudo-norm. The $\|\cdot\|_0$ is merely the cardinality of a signal's support set. For some value of s which depends on λ (2.8) is equivalent to the following constrained optimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{C}^n} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 \text{ subject to } \|\boldsymbol{x}\|_0 \leq s. \tag{2.9}$$

Since the ℓ_0 pseudo-norm is non-convex, the problems (2.8) and (2.9) are non-convex optimization problems. Yet another related problem is that of finding *sparsest vector*:

$$\min_{\boldsymbol{x}\in\mathbb{C}^n} \|\boldsymbol{x}\|_0 \text{ subject to } \|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2 \leq \eta, \tag{2.10}$$

which is an NP-Hard problem [14]. For the remainder of the chapter we will be concerned primarily with problem (2.10).

We conclude for that sparsity to be an attractive choice for a regularizer, the following properties must be satisfied:

1. **Well-posedness:** A priori, (2.10) is not a well posed problem. Can a sufficiently general class of sensing matrices be constructed which act as an injection on the space of sparse signals? This would certainly be a necessary condition to guarantee uniqueness of solutions to (2.10). Implicit in this is the number of rows of $\boldsymbol{A}$ or the number of measurements; how many measurements do we need to sufficiently distinguish between sparse signals?

2. **Tractable/Scalable Computational Complexity:** Can one design computationally efficient sparse approximation algorithms which scale better than the matrix inversion based solutions above?

2.3 RIP Matrices and Convex Relaxation

To address the first point, it turns out that there is a sufficiently general class of matrices which allow for compression/dimensionality reduction of high dimensional but sparse signals while simultaneously approximately preserving the signals ℓ_2 norm. To this end, we state the following definition from [1]:

Definition 2.3.1. If the entries of a matrix $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ are independent mean zero subgaussian random variables with variance 1 and some subgaussian parameters β, κ such that:

$$\Pr[|A_{j,k}| \geq t] \leq \beta \exp(-\kappa t^2), \quad \text{for all } t > 0, (j,k) \in [m] \times [n], \tag{2.11}$$

then $\boldsymbol{A}$ is called a subgaussian random matrix.

Subgaussian random matrices constitute a class of matrices which include Bernoulli matrices and (obviously) Gaussian matrices.

Soon we will see that with high probability these subgaussian random matrices satisfy a crucial property referred to as the *Restricted Isometry Property*:

Definition 2.3.2. The s-th restricted isometry constant $\delta_s = \delta_s(\boldsymbol{A})$ of a matrix $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ is the smallest $\delta \geq 0$ such that:

$$(1-\delta)\|\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{A}\boldsymbol{x}\|_2^2 \leq (1+\delta)\|\boldsymbol{x}\|_2^2 \tag{2.12}$$

for all sparse vectors $\boldsymbol{x} \in \mathbb{C}^n$. We say that $\boldsymbol{A}$ satisfies the RIP of order s with RIP constant $\delta_s(\boldsymbol{A})$.

The following crucial theorem from [1] establishes how much compression is possible using random subgaussian matrices:

Theorem 2.3.1. *Let $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ be a subgaussian random matrix. Then there exists a constant $C > 0$ (depending only on the subgaussian parameters (β, κ)) such that the restricted isometry constant of $\frac{1}{\sqrt{m}}\boldsymbol{A}$ satisfies $\delta_s \leq \delta$ with probability at least $(1-\epsilon)$ provided:*

$$m \geq C\delta^{-2}\left(s\ln(eN/s) + \ln(2\epsilon^{-1})\right). \tag{2.13}$$

Setting $\epsilon = 2\exp(-\delta^2 m/(2C))$ yields the condition:

$$m \geq 2C\delta^{-2}s\ln(eN/s), \tag{2.14}$$

which guarantees that $\delta_s \leq \delta$ with probability at least $1 - \exp(-\delta^2 m/(2C))$. Observe that this is a probabilistic guarantee. A random draw of a subgaussian matrix will with high probability satisfy the above, but there is no deterministic guarantee. In fact, it is still an open problem to deterministically construct such RIP sampling matrices $\boldsymbol{A}$ using the aforementioned sampling bounds on m.

We are now ready to address the second issue, namely that (2.10) is a non-convex problem. One of the major techniques in high dimensional data mining and machine learning has been to *relax non-convex regularizers/constraints to their tightest convex relaxation.*

Note that a basis for the space of sparse signals will be $\{\pm e_i\}_{i=1}^n$ where e_i denotes the standard basis vectors. Observe that the convex hull of these basis elements is precisely the ℓ_1 unit ball. We therefore see intuitively why the ℓ_1 norm is the tightest convex relaxation of the ℓ_0 pseudo-norm. For further details see [1].

Posing the convex relaxed problems, we have that (2.8) becomes relaxed to the *Basis Pursuit Denoising Problem* (BPDN):

$$\min_{\boldsymbol{x}\in\mathbb{C}^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1, \lambda > 0. \tag{2.15}$$

As before, for some value of s which depends on λ BPDN (2.15) is equivalent to the *LASSO* [15]:

$$\min_{\boldsymbol{x}\in\mathbb{C}^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 \text{ subject to } \|\boldsymbol{x}\|_1 \leq s. \tag{2.16}$$

For this section, we will focus on the convex relaxation of (2.10), referred to as the ℓ_1 minimization problem[1]:

$$\min_{\boldsymbol{x}\in\mathbb{C}^n} \|\boldsymbol{x}\|_1 \text{ subject to } \|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2 \leq \eta. \tag{2.17}$$

Using the RIP in conjunction with ℓ_1 minimization, from [1] we have the following performance guarantee:

Theorem 2.3.2. *Suppose that the 2s-th restricted isometry constant of the matrix $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ satisfies:*

$$\delta_{2s} < \frac{4}{\sqrt{41}}. \tag{2.18}$$

Then for any $\boldsymbol{x} \in \mathbb{C}^n$ and $\boldsymbol{y} \in \mathbb{C}^m$ with $\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2 \leq \eta$, a solution $\boldsymbol{x}^\#$ of (2.17) approximates the signal $\boldsymbol{x}$ with errors:

$$\begin{aligned}\|\boldsymbol{x}-\boldsymbol{x}^\#\|_1 &\leq C\sigma_s(\boldsymbol{x})_1 + D\sqrt{s}\eta,\\ \|\boldsymbol{x}-\boldsymbol{x}^\#\|_2 &\leq \frac{C}{\sqrt{s}}\sigma_s(\boldsymbol{x})_1 + D\eta,\end{aligned}$$

where the constants $C, D > 0$ only depend on δ_{2s} and $\sigma_s(\boldsymbol{x})_1 := \inf_{\|\boldsymbol{z}\|_0\leq s} \|\boldsymbol{x}-\boldsymbol{z}\|_1$.

Note that when $\boldsymbol{x}$ is actually s-sparse and there is no noise, i.e. $\eta = 0$ Theorem (2.3.2) guarantees exact sparse recovery.

The ℓ_1 minimization problem is a linear program and can be solved by convex optimization techniques. In general while these ℓ_1 constrained/regularized

[1]Note that BPDN, the LASSO and ℓ_1 minimization are all related to one another; for further details consult Section 3 of [1].

optimization problems are now convex and enjoy sparse recovery guarantees, these methods still scale poorly to high dimensional datasets. In what follows, we will consider greedy iterative methods, in particular one called Iterative Hard Thresholding. These methods have better scalability while still providing good sparse approximation guarantees.

2.4 Iterative Hard Thresholding

We will consider a problem of the form:

$$\min f(\boldsymbol{x}) \text{ subject to } \boldsymbol{x} \in \mathcal{S}, \tag{2.19}$$

where $\mathcal{S}$ is some structured solution space and $f(\boldsymbol{x})$ denotes a loss function which depends on $\boldsymbol{x}$ and the vector of linear samples $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{e}$, where $\boldsymbol{e}$ represents measurement noise and $\boldsymbol{A} \in \mathbb{C}^{m \times N}$ is a sampling matrix. For the case of best s-sparse approximation $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2$ and $\mathcal{S} = \{\boldsymbol{x} \in \mathbb{C}^N : \|\boldsymbol{x}\|_0 \leq s\}$, i.e. we are interested in identifying and approximating the top s-energetic components of an underlying signal $\boldsymbol{x}^*$ given a set of noisy linear measurements $\boldsymbol{y}$.

Greedy iterative algorithms solve problems like (2.19) by starting with some initial approximation $\boldsymbol{x}^0$ and iteratively making a set of computationally tractable but only locally optimal updates. For differentiable loss functions, this typically involves a gradient descent step. For compressed sensing problems, we have the smooth differentiable least square loss and computing these gradients typically only involve matrix-vector product operations, which scales

much more gracefully than matrix inversion or solving a linear program. That is why much research has been devoted to developing scalable and robust greedy iterative algorithms. Some of the more popular ones being OMP [16], CoSaMP [17] and Iterative Hard Thresholding (IHT) [12]. For the purposes of this thesis, we will now devote our attention to IHT.

Setting our initial approximation $\boldsymbol{x}^0 = \mathbf{0}$, IHT is the iteration:

$$\boldsymbol{x}^{n+1} = H_s(\boldsymbol{x}^n + \boldsymbol{A}^*(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^n)), \tag{2.20}$$

where H_s is the *hard thresholding operator* and at each step $n+1$ it outputs the best s-sparse approximation to $\boldsymbol{x}^n + \boldsymbol{A}^*(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^n)$ by projecting it onto $\mathcal{S}$. More specifically

$$H_s(\boldsymbol{x}) = \inf_{\|\boldsymbol{z}\|_0 \leq s} \|\boldsymbol{x} - \boldsymbol{z}\|_2. \tag{2.21}$$

Note that by plugging in for $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{e}$, we obtain the following approximation:

$$\boldsymbol{x}^n + \boldsymbol{A}^*(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^n) = \boldsymbol{A}^*\boldsymbol{A}\boldsymbol{x}^* + (\boldsymbol{I} - \boldsymbol{A}^*\boldsymbol{A})\boldsymbol{x}^n + \boldsymbol{A}^*\boldsymbol{e} \approx \boldsymbol{x}^* + \boldsymbol{\xi}^n.$$

Iterative greedy algorithms such as IHT exploit the RIP in the following manner. If the matrix $\boldsymbol{A}$ satisfies the RIP of order s, then $\boldsymbol{A}^*\boldsymbol{A}$ is a good enough approximation to the identity matrix on the space of s-sparse vectors so that applying $\boldsymbol{A}^*$ to the vector of noisy samples approximately yields the true signal $\boldsymbol{x}^*$ up to some "noise term"

$$\boldsymbol{A}^*\boldsymbol{y} = \boldsymbol{A}^*\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{A}^*\boldsymbol{e} \approx \boldsymbol{x}^* + \boldsymbol{\xi}.$$

Iterative greedy algorithms produce a dense signal approximation $\boldsymbol{z}^n \approx \boldsymbol{x}^* + \boldsymbol{\xi}^n$ at each stage n and they denoise this dense signal to output an s-sparse approximation. IHT denoises by applying the hard thresholding operator H_s, yielding $\boldsymbol{x}^{n+1}$, an s-sparse approximation to the true underlying signal $\boldsymbol{x}$. Roughly speaking, the denoising process of all of these iterative methods involves a projection onto the space of sparse vectors. The idea behind this is that while the noise term $\boldsymbol{\xi}^n$ is dense, its energy is assumed to be spread throughout all of its coordinates and as a result does not heavily contaminate the s-most significant coordinates of $\boldsymbol{x}^*$. Despite the fact that $\mathcal{S}$ is non-convex, it is very simple to compute projections onto this space: all that is required is sorting the entries of the signal by their magnitude and picking the top s entries. This fact combined with the above RIP intuition explains why greedy methods such as IHT can efficiently solve a non-convex problem.

In [12], the following performance guarantee was established[2]:

Theorem 2.4.1. *Let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$ denote a set of noisy observations where $\boldsymbol{x}$ is an arbitrary vector. Let $\boldsymbol{x}_s$ be an approximation to $\boldsymbol{x}$ with no more than s non-zero elements for which $\|\boldsymbol{x} - \boldsymbol{x}_s\|_2$ is minimal. If $\boldsymbol{A}$ has restricted isometry property with $\delta_{3s} < 1/\sqrt{32}$, then at iteration n, IHT as defined by (2.20) will recover an approximation $\boldsymbol{x}^n$ satisfying*

$$\|\boldsymbol{x} - \boldsymbol{x}^n\|_2 \leq 2^{-n}\|\boldsymbol{x}_s\|_2 + 6\widetilde{\epsilon}_s. \tag{2.22}$$

[2]We use different notation than that of the original authors Blumensath and Davies.

where

$$\widetilde{\epsilon}_s = \|\boldsymbol{x} - \boldsymbol{x}_s\|_2 + \frac{1}{\sqrt{s}}\|\boldsymbol{x} - \boldsymbol{x}_s\|_1 + \|\boldsymbol{e}\|_2. \tag{2.23}$$

In other words, IHT guarantees a linear convergence rate up to the *unrecoverable energy* $\widetilde{\epsilon}_s$. This $\widetilde{\epsilon}_s$ term is referred to as unrecoverable energy as it contains the measurement noise and energy terms which measure how well a signal $\boldsymbol{x}$ can be approximated by sparse signals.

2.5 Weighted Sparsity

In this section, all of the concepts and definitions are taken from [7]. For unstructured sparse recovery problems, the sparsity of a signal $\boldsymbol{x} \in \mathbb{C}^N$ is defined to be the cardinality of its support set, denoted as $\|\boldsymbol{x}\|_0$. More generally, we have a dictionary of atoms $\{a_i\}_{i=1}^N$ and for the unstructured case, each atom is given the weight $\omega_i = 1$ for all $i = 1, \ldots, N$. In this context, the sparsity of a signal can be viewed as the sum of the weights of the atoms in the support set. Following [7], given a dictionary $\{a_i\}_{i=1}^N$ and a corresponding set of weights $\{\omega_i\}_{i=1}^N, \omega_i \geq 1$ for $i = 1, \ldots, N$, we can define the *weighted* ℓ_0 *norm*:

$$\|\boldsymbol{x}\|_{\omega,0} = \sum_{j:x_j \neq 0} \omega_j^2.$$

Observe that the weighted sparsity of a vector $\boldsymbol{x}$ is at least as large as the unweighted sparsity of $\boldsymbol{x}$, i.e. $\|\boldsymbol{x}\|_{\omega,0} \geq \|\boldsymbol{x}\|_0$.

For any subset $S \subset [N]$, we may define the weighted cardinality of S

via:

$$\omega(S) := \sum_{j \in S} \omega_j^2.$$

In general, we also have the weighted ℓ_p spaces with norm:

$$\|\boldsymbol{x}\|_{\omega,p} = \sum_{j:x_j \neq 0} |x_j|^p \omega_j^{2-p}.$$

Using this generalized notion of sparsity allows us to pose the *best* (ω, s)*-sparse approximation problem:*

$$\text{minimize } \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 \text{ subject to } \|\boldsymbol{x}\|_{\omega,0} \leq s. \tag{2.24}$$

Given this generalized notion of sparsity, [7] defines a generalized notion of a map $\boldsymbol{A} : \mathbb{C}^N \to \mathbb{C}^m$ being an isometry on the space of weighted sparse vectors:

Definition 2.5.1. (Weighted restricted isometry constants) For $\boldsymbol{A} \in \mathbb{C}^{m \times N}$, weight parameter ω and $s \geq 1$, the weighted restricted isometry constant $\delta_{\omega,s}$ associated to $\boldsymbol{A}$ is the smallest number δ for which

$$(1 - \delta)\|\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{A}\boldsymbol{x}\|_2^2 \leq (1 + \delta)\|\boldsymbol{x}\|_2^2$$

holds for all $\boldsymbol{x} \in \mathbb{C}^N$ with $\|\boldsymbol{x}\|_{\omega,0} \leq s$. We say that a map $\boldsymbol{A}$ has the weighted restricted isometry property with respect to the weights ω (ω-RIP) if $\delta_{\omega,s}$ is small for s reasonably large compared to m.

Observe that for any positive number s, there exists a partition of s with distinct parts of maximal cardinality, i.e. an index set $\mathcal{I}$ with ω-weighted

cardinality s with the largest number of non-zero atoms. Let $P_\omega(s)$ denote this maximal term

$$P_\omega(s) := \max_{\omega(\mathcal{J})\leq s} |\mathcal{J}|.$$

Clearly if $\boldsymbol{A}$ satisfies RIP of order $P_\omega(s)$, then $\boldsymbol{A}$ will also satisfy the ω-RIP of order s. However the converse does not hold. Not only do weighted sparse signals have a constraint on the cardinality of their support sets, they can also have a constraint on the maximal atom which can be present in their support sets. Take for example the weights defined by $\omega(j) = \sqrt{j},\ j = 1, \ldots, N$. An (ω, s)-sparse signal cannot have any atom with index higher than $\lceil s^{1/2} \rceil$ supported. If $\boldsymbol{A}$ were to satisfy the ω-RIP of order s, then the ω-RIP alone does not guarantee that $\boldsymbol{A}$ preserves the geometry of heavy-tailed signals, no matter how sparse they may be in the unweighted sense. We conclude that the ω-RIP is in general a weaker isometry condition than the RIP and the primary reason being the existence of heavy tailed signals.

In [7] a heuristic justification is given that for weights ω satisfying $\omega(j) = j^{\alpha/2}$, with high probability an $m \times N$ i.i.d. subgaussian random matrix satisfies the ω-RIP once

$$m = \mathcal{O}\left(\alpha^{1/\alpha - 1} s^{1/(\alpha+1)} \log(s)\right).$$

Note that fewer measurements are required than in the unweighted case, which has the lower bound of $m = \mathcal{O}(s \log(N/s))$ measurements.

The following properties of RIP matrices carry over immediately to ω-RIP matrices.

Lemma 2.5.1. *Let $\boldsymbol{I}$ denote the $N \times N$ identity matrix. Given a set of weights $\omega \in \mathbb{C}^N$, vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^N$, $\boldsymbol{y} \in \mathbb{C}^m$ and an index set $S \subseteq [N]$,*

$$
\begin{aligned}
|\langle \boldsymbol{u}, (\boldsymbol{I} - \boldsymbol{A}^*\boldsymbol{A})\boldsymbol{v}\rangle| &\leq \delta_{\omega,t}\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2 && if\ \|\mathrm{supp}(\boldsymbol{u}) \cup \mathrm{supp}(\boldsymbol{v})\|_{\omega,0} \leq t, \\
\|((\boldsymbol{I} - \boldsymbol{A}^*\boldsymbol{A})\boldsymbol{v})_S\|_2 &\leq \delta_{\omega,t}\|\boldsymbol{v}\|_2 && if\ \|S \cup \mathrm{supp}(\boldsymbol{v})\|_{\omega,0} \leq t, \\
\|(\boldsymbol{A}^*\boldsymbol{y})_S\|_2 &\leq \sqrt{1+\delta_{\omega,s}}\|\boldsymbol{y}\|_2 && if\ \|S\|_{\omega,0} \leq s.
\end{aligned}
$$

Proof. The proofs follow immediately from their unweighted counterparts where one employs the ω-RIP instead of the RIP. See [1] for full proofs. □

2.6 Iterative Hard Weighted Thresholding: In Theory

2.6.1 Extension to the Weighted Case

Observe that one can equivalently view IHT as a projected gradient descent algorithm with constant step size equal to 1. Once IHT is viewed in this manner, the modification we make to extend IHT to solve (2.24) is quite natural: we still perform a constant step size gradient descent step at each iterate, however instead of projecting onto the space of s-sparse vectors, we project onto the space of weighted sparse vectors $\mathcal{S}_{\omega,s} = \{\boldsymbol{x} : \|\boldsymbol{x}\|_{\omega,0} \leq s\}$. This algorithm will be referred to as Iterative Hard Weighted Thresholding (IHWT) and it is given by the following iteration

$$\boldsymbol{x}^{n+1} = H_{\omega,s}(\boldsymbol{x}^n + \boldsymbol{A}^*(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^n)), \tag{2.25}$$

where $H_{\omega,s}$ is the *hard weighted thresholding operator* and it computes projections onto the space of weighted sparse vectors $\mathcal{S}_{\omega,s}$

$$H_{\omega,s}(\boldsymbol{x}) = \inf_{\|\boldsymbol{z}\|_{\omega,0} \leq s} \|\boldsymbol{x} - \boldsymbol{z}\|_2. \tag{2.26}$$

Computing the projection $H_{\omega,s}(\boldsymbol{x})$ is not as straightforward as computing $H_s(\boldsymbol{x})$. In particular, sorting the signal by the magnitude of its entries and then thresholding does not produce the best (ω, s)-sparse approximation. To see why, consider the simple example where $N = 3, \omega = [1, \sqrt{2}, \sqrt{3}]$ and take the signal $\boldsymbol{x} = [9, 9, 10]$. For $s = 3$, by sorting and thresholding, we obtain the following weighted 3 sparse approximation $\boldsymbol{x}^* = [0, 0, 10]$ and $\|\boldsymbol{x} - \boldsymbol{x}^*\|_2 = 9\sqrt{2}$. However, the other 3 sparse approximation $\widehat{\boldsymbol{x}} = [9, 9, 0]$ is in fact a more accurate weighted 3 sparse approximation as $\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_2 = 10 < 9\sqrt{2}$.

Therefore unlike the unweighted case, computing the best weighted s-sparse approximation consists of a combinatorial search. To illustrate the difficulty of executing this search, consider the case of a weight parameter ω given by $\omega(j) = \sqrt{j}$ for $j = 1, \ldots, N$. In this case, computing all the possible index sets of weighted cardinality s is equivalent to computing all the partitions of s consisting of unique parts. With the square root weight parameter, Wolfram Mathematica [18] computes that there are 444,794 possible subsets of weighted sparsity $s = 100$ and it computes that there are 8,635,565,795,744,155,161,506 support sets of size $s = 1000$.

Despite this intractability, in the next subsection we derive theoretical guarantees for the IHWT algorithm and in Section 2.7, we will explore the

empirical performance of a computationally efficient surrogate to approximate the projection onto $\mathcal{S}_{\omega,s}$.

2.6.2 Performance Guarantees

Throughout this subsection, we will employ the following notation:

1. $\boldsymbol{x}_s = H_{\omega,s}(\boldsymbol{x})$ as defined by (2.26),
2. $\boldsymbol{r}^n = \boldsymbol{x}_s - \boldsymbol{x}^n$,
3. $S = \text{supp}(\boldsymbol{x}_s)$,
4. $\boldsymbol{x}_{\overline{S}} = \boldsymbol{x} - \boldsymbol{x}_s$,
5. $S^n = \text{supp}(\boldsymbol{x}^n)$,
6. $T^n = S \cup S^n$,
7. $\boldsymbol{a}^{n+1} = \boldsymbol{x}^n + \boldsymbol{A}^*(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^n) = \boldsymbol{x}^n + \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{e} - \boldsymbol{A}\boldsymbol{x}^n) = \boldsymbol{x}^n + \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{x}_s + \boldsymbol{A}\boldsymbol{x}_{\overline{S}} + \boldsymbol{e} - \boldsymbol{A}\boldsymbol{x}^n)$.

2.6.2.1 Performance Guarantees: Convergence to a Neighborhood

Here we derive performance guarantees which establish that IHWT will converge to a neighborhood of the best (ω, s)-sparse approximation with a linear convergence rate. The size of the neighborhood is dependent on how well the true signal $\boldsymbol{x}$ is approximated by $\boldsymbol{x}_s$.

To begin we focus our attention on an intermediate, yet more general error bound for IHT:

Theorem 2.6.1. *Let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$ denote a set of noisy observations where $\boldsymbol{x}$ is an arbitrary vector. Let $\boldsymbol{x}_s$ be an approximation to s with no more than s non-zero elements for which $\|\boldsymbol{x}-\boldsymbol{x}_s\|_2$ is minimal. If $\boldsymbol{A}$ has restricted isometry property with $\delta_{3s} < 1/\sqrt{32}$, then at iteration n, IHT as defined by (2.20) will recover an approximation $\boldsymbol{x}^n$ satisfying*

$$\|\boldsymbol{x} - \boldsymbol{x}^n\|_2 \leq 2^{-n}\|\boldsymbol{x}_s\|_2 + \|\boldsymbol{x} - \boldsymbol{x}_s\|_2 + 4.34\|\boldsymbol{A}\boldsymbol{x}_{\overline{S}} + \boldsymbol{e}\|_2. \tag{2.27}$$

To pass from (2.27) to (2.22)–(2.23), Blumensath and Davies used the following energy bound for RIP matrices from [17]:

Proposition 2.6.2. *Suppose that $\boldsymbol{A}$ verifies the upper inequality*

$$\|\boldsymbol{A}\boldsymbol{x}\|_2 \leq \sqrt{1+\delta_s}\|\boldsymbol{x}\|_2, \text{ when } \|\boldsymbol{x}\|_0 \leq s.$$

Then, for every signal $\boldsymbol{x}$,

$$\|\boldsymbol{A}\boldsymbol{x}\|_2 \leq \sqrt{1+\delta_s}\left[\|\boldsymbol{x}\|_2 + \frac{1}{\sqrt{s}}\|\boldsymbol{x}\|_1\right]. \tag{2.28}$$

Applying (2.28) to $\boldsymbol{A}\boldsymbol{x}_{\overline{S}}$ in (2.27) yields (2.22)–(2.23).

The proof of Proposition 2.6.2 boils down to establishing an inclusion of polar spaces: $S^\circ \subset K^\circ$. S° is equipped with the following norm

$$\|\boldsymbol{u}\|_{S^\circ} = \max_{|I|\leq r} \|\boldsymbol{u}_I\|_2.$$

The proof proceeds by considering any element $\boldsymbol{u}$ of the unit ball in S°. We decompose $\boldsymbol{u}$ into two components: $\boldsymbol{u}_S$ and $\boldsymbol{u}_{\overline{S}}$ where $\boldsymbol{u}_S$ represents the best s-sparse approximation to $\boldsymbol{u}$ in the ℓ_2 norm. As S contains the s most energetic

atoms, this implies that the set $\overline{S}$ contains atoms whose energy must lie under a certain threshold: the bound $\|\boldsymbol{u}_{\overline{S}}\|_\infty \leq \frac{1}{\sqrt{s}}$ is easily obtained. In other words, the following decomposition is obtained:

$$\boldsymbol{u} = \boldsymbol{u}_S + \boldsymbol{u}_{\overline{S}} \in B_2 + \frac{1}{\sqrt{s}} B_\infty,$$

and the space on the right hand side is exactly the space K°. For further details, consult [17], in this chapter we will only be concerned with this particular aspect of their proof.

This sort of decomposition does not hold for the weighted case. Consider the example in which the weight vector ω is such that $\omega(j) = \sqrt{j}$. As mentioned before, with such a weight vector ω, any s sparse signal cannot have any atom of index higher than $\lceil s^{1/2} \rceil$ supported. Therefore, taking the best (ω, s)-sparse approximation to a signal does not constrain the ℓ_∞ norm of the signal on the complement $\overline{S}$. As a result of this, Proposition 2.6.2 does not extend to the weighted case and an alternative method will be needed to bound the energy of $\boldsymbol{A}\boldsymbol{x}_{\overline{S}}$. Here we see a key difference between unweighted sparsity and weighted sparsity: significant amounts of energy can be concentrated in the tail $\boldsymbol{x} - H_{\omega,s}(\boldsymbol{x})$. More specifically we see that certain weight vectors can yield the process of taking the best (ω, s)-sparse approximation to be an operation which is inherently local as it may restrict the analysis to lie on a subset of low weight atoms and the higher weight atoms are completely ignored.

An alternative method of bounding the term $\|\boldsymbol{A}\boldsymbol{x}\|_2$ involves a different

type of decomposition. Suppose $\boldsymbol{A}$ satisfies the RIP of order s with RIP constant δ_s. Let $\{S_i\}_{i=1}^p$ be a partition of $[N]$ into s-sparse blocks: each S_i satisfies: for all $i \neq j, S_i \cap S_j = \emptyset$ and $\text{card}(S_i) \leq s$. We may apply the RIP to each S_i block to obtain the following bound:

$$\begin{aligned}\|\boldsymbol{A}\boldsymbol{x}\|_2 = \|\boldsymbol{A}(\sum_i \boldsymbol{x}_{S_i})\|_2 &\leq \sum_i \|\boldsymbol{A}\boldsymbol{x}_{S_i}\|_2 \\ &\leq \sqrt{1+\delta_s}\sum_i \|\boldsymbol{x}_{S_i}\|_2 \\ &\leq \sqrt{(1+\delta_s)/s}\sum_i \|\boldsymbol{x}_{S_i}\|_1 = \sqrt{(1+\delta_s)/s}\|\boldsymbol{x}\|_1\end{aligned}$$

This sort of argument in general will not extend to the weighted case. Depending on the weight vector ω such a decomposition of an arbitrary signal into a collection of disjoint s-sparse blocks may not even be possible.

Therefore, the performance guarantee given in Theorem 2.4.1 does not directly extend to the weighted case. The more general guarantee of Theorem 2.6.1 however, does extend, and we present the proof:

Theorem 2.6.3. *Let $\omega \in \mathbb{C}^N$ denote a weight vector with $\omega(i) \geq 1$ for all $i = 1, \ldots, N$. Let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$ denote a set of noisy observations where $\boldsymbol{x}$ is an arbitrary vector. If $\boldsymbol{A}$ has weighted restricted isometry property with $\delta_{\omega,3s} < 1/\sqrt{32}$, then at iteration n, IHWT as defined by (2.25) will recover an approximation $\boldsymbol{x}^n$ satisfying*

$$\|\boldsymbol{x} - \boldsymbol{x}^n\|_2 \leq 2^{-n}\|\boldsymbol{x}_s\|_2 + \|\boldsymbol{x} - \boldsymbol{x}_s\|_2 + 4.34\|\boldsymbol{A}\boldsymbol{x}_{\overline{S}} + \boldsymbol{e}\|_2. \tag{2.29}$$

Proof. We follow the proof presented in [12]. By the triangle inequality we

have that:

$$\|\boldsymbol{x} - \boldsymbol{x}^{n+1}\|_2 \leq \|\boldsymbol{x} - \boldsymbol{x}_s\|_2 + \|\boldsymbol{x}^{n+1} - \boldsymbol{x}_s\|_2. \tag{2.30}$$

We focus on the term $\|\boldsymbol{x}^{n+1} - \boldsymbol{x}_s\|_2$. This term is supported on T^{n+1} and we may therefore restrict our analysis to this index set. By the triangle inequality we have:

$$\|\boldsymbol{x}_s - \boldsymbol{x}^{n+1}\|_2 \leq \|(\boldsymbol{x}_s)_{T^{n+1}} - \boldsymbol{a}^{n+1}_{T^{n+1}}\|_2 + \|\boldsymbol{x}^{n+1}_{T^{n+1}} - \boldsymbol{a}^{n+1}_{T^{n+1}}\|_2$$

By definition of the thresholding operator $H_{\omega,s}$, the signal $\boldsymbol{x}^{n+1}$ is the best weighted s sparse approximation to $\boldsymbol{a}^{n+1}$. In particular, $\boldsymbol{x}^{n+1}$ is a better weighted s sparse approximation to $\boldsymbol{a}^{n+1}$ than $\boldsymbol{x}_s$. We therefore obtain the inequality:

$$\|\boldsymbol{x}_s - \boldsymbol{x}^{n+1}\|_2 \leq 2\|(\boldsymbol{x}_s)_{T^{n+1}} - \boldsymbol{a}^{n+1}_{T^{n+1}}\|_2.$$

Expanding the term $\boldsymbol{a}^{n+1}$:

$$\begin{aligned}
\|\boldsymbol{x}_s - \boldsymbol{x}^{n+1}\|_2 &\leq 2\|(\boldsymbol{x}_s)_{T^{n+1}} - \boldsymbol{x}^n_{T^{n+1}} - \boldsymbol{A}^*_{T^{n+1}}\boldsymbol{A}\boldsymbol{r}^n + \boldsymbol{A}^*_{T^{n+1}}\boldsymbol{A}\boldsymbol{x}_{\overline{S}} + \boldsymbol{A}^*_{T^{n+1}}\boldsymbol{e}\|_2 \\
&\leq 2\|\boldsymbol{r}^n_{T^{n+1}} - \boldsymbol{A}^*_{T^{n+1}}\boldsymbol{A}\boldsymbol{r}^n\|_2 + 2\|\boldsymbol{A}^*_{T^{n+1}}(\boldsymbol{A}\boldsymbol{x}_{\overline{S}} + \boldsymbol{e})\|_2 \\
&= 2\|(\boldsymbol{I} - \boldsymbol{A}^*_{T^{n+1}}\boldsymbol{A}_{T^{n+1}})\boldsymbol{r}^n_{T^{n+1}} - \boldsymbol{A}^*_{T^{n+1}}\boldsymbol{A}_{T^n \setminus T^{n+1}}\boldsymbol{r}^n_{T^n \setminus T^{n+1}}\|_2 \\
&\qquad + 2\|\boldsymbol{A}^*_{T^{n+1}}(\boldsymbol{A}\boldsymbol{x}_{\overline{S}} + \boldsymbol{e})\|_2 \\
&\leq 2\|(\boldsymbol{I} - \boldsymbol{A}^*_{T^{n+1}}\boldsymbol{A}_{T^{n+1}})\boldsymbol{r}^n_{T^{n+1}}\|_2 + 2\|\boldsymbol{A}^*_{T^{n+1}}\boldsymbol{A}_{T^n \setminus T^{n+1}}\boldsymbol{r}^n_{T^n \setminus T^{n+1}}\|_2 \\
&\qquad + 2\|\boldsymbol{A}^*_{T^{n+1}}(\boldsymbol{A}\boldsymbol{x}_{\overline{S}} + \boldsymbol{e})\|_2
\end{aligned}$$

Note that $T^n \setminus T^{n+1}$ is disjoint from T^{n+1} and $\|T^n \cup T^{n+1}\|_{\omega,0} \leq 3s$. Applying

the RIP bounds from 2.5.1:

$$\begin{aligned}\|\boldsymbol{r}^{n+1}\|_2 &\leq 2\delta_{\omega,2s}\|\boldsymbol{r}^n_{T^{n+1}}\|_2 + 2\delta_{\omega,3s}\|\boldsymbol{r}^n_{T^n\setminus T^{n+1}}\|_2 + 2\sqrt{1+\delta_{\omega,2s}}\|\boldsymbol{A}\boldsymbol{x}_{\overline{S}} + \boldsymbol{e}\|_2 \\ &\leq 2\delta_{\omega,3s}\left(\|\boldsymbol{r}^n_{T^{n+1}}\|_2 + \|\boldsymbol{r}^n_{T^n\setminus T^{n+1}}\|_2\right) + 2\sqrt{1+\delta_{\omega,3s}}\|\boldsymbol{A}\boldsymbol{x}_{\overline{S}} + \boldsymbol{e}\|_2 \\ &\leq \sqrt{8}\delta_{\omega,3s}\|\boldsymbol{r}^n\|_2 + 2\sqrt{1+\delta_{\omega,3s}}\|\boldsymbol{A}\boldsymbol{x}_{\overline{S}} + \boldsymbol{e}\|_2.\end{aligned}$$

If we have that $\delta_{\omega,3s} < \frac{1}{\sqrt{32}}$, then

$$\|\boldsymbol{r}^{n+1}\|_2 \leq 0.5\|\boldsymbol{r}^n\|_2 + 2.17\|\boldsymbol{A}\boldsymbol{x}_{\overline{S}} + e\|_2.$$

Iterating this relationship and using the fact that $\sum_{i=0}^{\infty} 2^{-i} = 2$, we obtain the bound:

$$\|\boldsymbol{r}^n\|_2 < 2^{-n}\|\boldsymbol{x}_s\|_2 + 4.34\|\boldsymbol{A}\boldsymbol{x}_{\overline{S}} + \boldsymbol{e}\|_2. \tag{2.31}$$

Combining (2.31) with (2.30) proves the desired claim. □

Note that the proof has two main components: the hard thresholding operator produces $\boldsymbol{x}^{n+1}$, a superior sparse approximation to the gradient descent update $\boldsymbol{a}^{n+1}$ than $\boldsymbol{x}_s$ and applying the RIP. Moreover, the proof never requires any details of the projection or even the space we are projecting onto, unlike the proof of Proposition 2.6.2, which uses special properties of the projection onto the space of unweighted s sparse signals.

The existence of weights which are known to produce signals with heavy tails is the main blockage to the extension of some more detailed performance guarantees, like that of Theorem 2.6.1. In the two cases before, one notices that the existence of heavy tailed signals prevented a decomposition of a signal

amenable to further analysis. However, for certain *bounded* weight parameters, for arbitrary signals $\boldsymbol{x}$, one may obtain a modified bound on $\|\boldsymbol{A}\boldsymbol{x}\|_2$ in terms of *weighted norms*. Indeed we obtain the following weighted analogue of Proposition 2.6.2:

Proposition 2.6.4. *Consider a sparsity level s and a weight parameter ω satisfying $s \geq 2\|\omega\|_\infty^2$. If $\boldsymbol{A}$ satisfies the ω-RIP of order s with RIP constant $\delta_{\omega,s}$, then the following inequality holds for any arbitrary signal $\boldsymbol{x}$:*

$$\|\boldsymbol{A}\boldsymbol{x}\|_2 \leq \sqrt{1+\delta_{\omega,s}}\left(\|\boldsymbol{x}\|_2 + \frac{2}{\sqrt{s}}\|\boldsymbol{x}\|_{\omega,1}\right) \tag{2.32}$$

Proof. The proof employs the same strategy used in the proof of Theorem 4.5 in [7]. Let $\boldsymbol{x} \in \mathbb{C}^N$. We will partition $[N]$ into weighted s sparse blocks $S_1, \ldots, S_p$ for some index p with each block satisfying $s - \|\omega\|_\infty^2 \leq \omega(S_l) \leq s$. Furthermore, we assume that the blocks S_i are formed according to a *non-increasing rearrangement* of $\boldsymbol{x}$ with respect to the weights, i.e.

$$|x_j|\omega_j^{-1} \leq |x_k|\omega_k^{-1} \text{ for all } j \in S_l \text{ and for all } k \in S_{l-1}, l \geq 2. \tag{2.33}$$

For any $k \in S_l$, set $\alpha_k = (\sum_{j\in S_l}\omega_j^2)^{-1}\omega_k^2 \leq (s - \|\omega\|_\infty^2)^{-1}\omega_k^2$ by hypothesis. Notice that $\sum_{k\in S_l}\alpha_k = 1$. For $l \geq 2$ then:

$$\begin{aligned}
|x_j|\omega_j^{-1} &\leq \sum_{k\in S_{l-1}} \alpha_k |x_k|\omega_k^{-1} \text{ for any } j \in S_l && (2.34)\\
&\leq (s - \|\omega\|_\infty^2)^{-1} \sum_{k\in S_{l-1}} |x_k|\omega_k^{-1}\omega_k^2 && (2.35)\\
&= (s - \|\omega\|_\infty^2)^{-1} \sum_{k\in S_{l-1}} |x_k|\omega_k && (2.36)\\
&= (s - \|\omega\|_\infty^2)^{-1}\|\boldsymbol{x}_{S_{l-1}}\|_{\omega,1}. && (2.37)
\end{aligned}$$

where (2.34) holds by non-increasing rearrangement and convexity and (2.35) holds by hypothesis. Therefore, by the Cauchy-Schwarz inequality, we obtain:

$$\|\boldsymbol{x}_{S_l}\|_2 \leq \frac{\sqrt{s}}{s - \|\omega\|_\infty^2}\|\boldsymbol{x}_{S_{l-1}}\|_{\omega,1} \leq \frac{2}{\sqrt{s}}\|\boldsymbol{x}_{S_{l-1}}\|_{\omega,1} \text{ for } l \geq 2.$$

For $\|\boldsymbol{A}\boldsymbol{x}\|_2$, we obtain the following estimate:

$$\begin{aligned}
\|\boldsymbol{A}\boldsymbol{x}\|_2 &\leq \sum_{i=1}^{p} \|\boldsymbol{A}\boldsymbol{x}_{S_i}\|_2 \\
&\leq \sqrt{1+\delta_{\omega,s}} \sum_{i=1}^{p} \|\boldsymbol{x}_{S_i}\|_2 \\
&= \sqrt{1+\delta_{\omega,s}} \left(\|\boldsymbol{x}_{S_1}\|_2 + \sum_{i=2}^{p} \|\boldsymbol{x}_{S_i}\|_2 \right) \\
&\leq \sqrt{1+\delta_{\omega,s}} \left(\|\boldsymbol{x}_{S_1}\|_2 + \frac{2}{\sqrt{s}} \sum_{i=1}^{p-1} \|\boldsymbol{x}_{S_i}\|_{\omega,1} \right) \\
&\leq \sqrt{1+\delta_{\omega,s}} \left(\|\boldsymbol{x}\|_2 + \frac{2}{\sqrt{s}} \|\boldsymbol{x}\|_{\omega,1} \right).
\end{aligned}$$

□

Applying 2.6.4 to $\|\boldsymbol{A}\boldsymbol{x}_{\overline{S}}\|_2$ immediately yields the following performance guarantee:

Theorem 2.6.5. *For sparsity level s, let $\omega \in \mathbb{C}^N$ denote a weight vector with $\omega(i) \geq 1$ for all $i = 1, \ldots, N$ satisfying $s \geq 2\|\omega\|_\infty^2$. Let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$ denote a set of noisy observations where $\boldsymbol{x}$ is an arbitrary vector. If $\boldsymbol{A}$ has weighted restricted isometry property with $\delta_{\omega,3s} < 1/\sqrt{32}$, then at iteration n, IHWT as*

defined by (2.25) *will recover an approximation* $\boldsymbol{x}^n$ *satisfying*

$$\|\boldsymbol{x} - \boldsymbol{x}^n\|_2 \leq 2^{-n}\|\boldsymbol{x}_s\|_2 + 6\left(\|\boldsymbol{x} - \boldsymbol{x}_s\|_2 + \frac{2}{\sqrt{s}}\|\boldsymbol{x} - \boldsymbol{x}_s\|_{\omega,1} + \|\boldsymbol{e}\|_2\right). \quad (2.38)$$

This result bears a striking resemblance to Theorem 2.4.1 except that it is in terms of the weighted ℓ_1 norm as opposed to the unweighted ℓ_1 norm.

2.6.2.2 Performance Guarantees: Contraction

For an arbitrary, possibly dense signal $\boldsymbol{x}$, the performance guarantees presented above do not guarantee the convergence of IHT/IHWT, but rather they guarantee that if the sampling matrix $\boldsymbol{A} \in \mathbb{C}^{m\times N}$ satisfies the RIP of order $3s$ then the iterates are guaranteed to converge to a *neighborhood* of the true best s-sparse approximation. In [19], alternative guarantees are derived under an alternative assumption on $\boldsymbol{A}$, namely that $\|\boldsymbol{A}\|_2 < 1$. In particular, we focus on the guarantee that if $\boldsymbol{A}$ satisfies the spectral bound $\|\boldsymbol{A}\|_2 < 1$, then the sequence of IHT iterates $(\boldsymbol{x}_n)$ is a contractive sequence.

Note that if $\boldsymbol{A}$ satisfies the RIP of order $3s$, then by applying $\boldsymbol{A}$ to the canonical Euclidean basis vectors $\{e_i\}_{i=1}^N$ it follows that the ℓ_2 norm of the columns of $\boldsymbol{A}$ must satisfy:

$$1 - \delta_{3s} \leq \|A_j\|_2 \leq 1 + \delta_{3s}, \text{ for } j = 1, \ldots, N. \quad (2.39)$$

On the other hand, the spectral norm of a linear map can equivalently be interpreted as an operator norm: $\|\boldsymbol{A}\|_2 = \sup_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_2$. As a consequence:

$$\|\boldsymbol{A}\|_2 = \sup_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_2 \geq \max_{\boldsymbol{e}_i, i=1,\ldots,N} \|\boldsymbol{A}\boldsymbol{e}_i\|_2 = \max_{i=1,\ldots,N} \|A_i\|_2.$$

Therefore if $\boldsymbol{A}$ satisfies the RIP condition, it could be true that $\max_{i=1,\ldots,N} \|A_i\|_2 > 1$ by (2.39). In this manner, the RIP condition is in general not compatible with the spectral condition $\|\boldsymbol{A}\|_2 < 1$.

Observe that if the spectral norm of $\boldsymbol{A}$ is bounded above by 1, then the loss function $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$ is *majorized* by the following surrogate objective function:

$$g(\boldsymbol{x}, \boldsymbol{z}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 - \|\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{z})\|_2^2 + \|\boldsymbol{x} - \boldsymbol{z}\|_2^2. \tag{2.40}$$

Because $g(\boldsymbol{x}, \boldsymbol{x}) = f(\boldsymbol{x})$, optimizing $g(\boldsymbol{x}, \boldsymbol{x})$ will decrease the objective function $f(\boldsymbol{x})$. This is known as Lange's *Majorization Minimization* (MM) Method [20].

Viewing $\boldsymbol{z}$ as fixed, we may decouple the coordinates x_i:

$$g(\boldsymbol{x}, \boldsymbol{z}) \propto \sum_i x_i^2 - 2x_i(z_i + A_i^*\boldsymbol{y} - A_i^*\boldsymbol{A}\boldsymbol{z}). \tag{2.41}$$

Ignoring the sparsity constraint on $\boldsymbol{x}$, minimizing (2.41) we obtain the unconstrained minima $\boldsymbol{x}^*$ given by:

$$x_i^* = z_i + A_i^*\boldsymbol{y} - A_i^*\boldsymbol{A}\boldsymbol{z}.$$

We then have that:

$$g(\boldsymbol{x}^*, \boldsymbol{z}) \propto \sum_i x_i^{*2} - 2x_i^*(z_i + A_i^*\boldsymbol{y} - A_i^*\boldsymbol{A}\boldsymbol{z}) = \sum_i -x_i^{*2}.$$

Therefore the *s-sparse constrained minimum* of the majorizing surrogate g is given by hard thresholding $\boldsymbol{x}^*$ by choosing the largest s components in magnitude. Clearly the above analysis holds for weighted sparse approximations

as well. We therefore conclude that both the IHT and IHWT iterates share the property that the sparsity constrained minimizer of $g(\boldsymbol{x}, \boldsymbol{x}^n)$ is given by $\boldsymbol{x} = \boldsymbol{x}^{n+1}$.

The following lemma establishes that IHWT makes progress at each iterate.

Lemma 2.6.6. *Assume that $\|\boldsymbol{A}\|_2 < 1$ and let $(\boldsymbol{x}^n)$ denote the IHWT iterates defined by (2.25). Then the sequences $(f(\boldsymbol{x}^n))$ and $(g(\boldsymbol{x}^{n+1}, \boldsymbol{x}^n))$ are non-increasing.*

Proof. We have the following sequence of inequalities:

$$\begin{aligned} f(\boldsymbol{x}^{n+1}) &\leq f(\boldsymbol{x}^{n+1}) + \|\boldsymbol{x}^{n+1} - \boldsymbol{x}^n\|_2^2 - \|A(\boldsymbol{x}^{n+1} - \boldsymbol{x}^n)\|_2^2 \\ &= g(\boldsymbol{x}^{n+1}, \boldsymbol{x}^n) \\ &\leq g(\boldsymbol{x}^n, \boldsymbol{x}^n) \\ &= f(\boldsymbol{x}^n) \\ &\leq f(\boldsymbol{x}^n) + \|\boldsymbol{x}^n - \boldsymbol{x}^{n-1}\|_2^2 - \|A(\boldsymbol{x}^n - \boldsymbol{x}^{n-1})\|_2^2 \\ &= g(\boldsymbol{x}^n, \boldsymbol{x}^{n-1}). \end{aligned}$$

□

Next we present the following lemma which states that the IHWT iterates contract.

Lemma 2.6.7. *If the sensing matrix $\boldsymbol{A}$ satisfies $\|A\|_2^2 \leq 1 - c < 1$ for some positive $c \in (0, 1)$, then for the IHWT iterates $(\boldsymbol{x}^n)$ the following limit holds: $\lim_{n\to\infty} \|\boldsymbol{x}^{n+1} - \boldsymbol{x}^n\|_2^2 = 0$.*

Proof. By the spectral bound:

$$\|\boldsymbol{A}(\boldsymbol{x}^{n+1} - \boldsymbol{x}^n)\|_2^2 \leq (1-c)\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^n\|_2^2.$$

Rearranging terms

$$\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^n\|_2^2 \leq \frac{1}{c}\left[\|\boldsymbol{x}^{n+1} - \boldsymbol{x}^n\|_2^2 - \|A(\boldsymbol{x}^{n+1} - \boldsymbol{x}^n)\|_2^2\right].$$

We define the sequence of partial sums (s_n) by $s_n = \sum_{i=0}^{n} \|\boldsymbol{x}^{i+1} - \boldsymbol{x}^i\|_2^2$. Clearly the sequence (s_n) is monotonically increasing. If we can show that the sequence (s_n) is also bounded, then (s_n) is a convergent sequence. Let k be any arbitrary index. We then obtain the following sequence of inequalities:

$$\begin{aligned}
s_k = \sum_{i=0}^{k} \|\boldsymbol{x}^{i+1} - \boldsymbol{x}^i\|_2^2 &\leq \frac{1}{c}\sum_{i=0}^{k}\left(\|\boldsymbol{x}^{i+1} - \boldsymbol{x}^i\|_2^2 - \|A(\boldsymbol{x}^{i+1} - \boldsymbol{x}^i)\|_2^2\right) && (2.42)\\
&= \frac{1}{c}\sum_{i=0}^{k} g(\boldsymbol{x}^{i+1}, \boldsymbol{x}^i) - \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^{i+1}\|_2^2 && (2.43)\\
&\leq \frac{1}{c}\sum_{i=0}^{k} g(\boldsymbol{x}^i, \boldsymbol{x}^i) - f(\boldsymbol{x}^{i+1}) && (2.44)\\
&= \frac{1}{c}\sum_{i=0}^{k} f(\boldsymbol{x}^i) - f(\boldsymbol{x}^{i+1}) && (2.45)\\
&= \frac{1}{c}(f(\boldsymbol{x}^0) - f(\boldsymbol{x}^{k+1})) && (2.46)\\
&\leq \frac{1}{c} f(\boldsymbol{x}^0) && (2.47)
\end{aligned}$$

where (2.44) follows from the next IHWT iterate $\boldsymbol{x}^{i+1}$ being a minimizer of $g(\boldsymbol{x}, \boldsymbol{x}^i)$.

Therefore $\{s_n\}$ is a convergent sequence. As the sequence of partial sums converges, the infinite sum $\sum_{i=0}^{\infty} \|\boldsymbol{x}^{i+1} - \boldsymbol{x}^i\|_2^2 < \infty$ and as a result $\lim_{n\to\infty} \|\boldsymbol{x}^{n+1} - \boldsymbol{x}^n\|_2^2 = 0$. □

2.7 Iterative Hard Weighted Thresholding: In Practice

2.7.1 Choosing the weights

Before delving into numerical experiments, we pause for a moment and focus on the overall setup of performing signal analysis in practice. Note that we have ignored the preprocessing required to properly select the weight parameter ω. In reality, this may require either significant domain knowledge (hand crafted) or the application of a learning algorithm to rank the atoms and assign weights (learned). If $N \gg 1$, it is not feasible to expect a human expert to assign weights to each of these atoms and instead we may assign weights to blocks of atoms. While this may be effective, the overall structure of the signals may not be fully captured in such a model. It is an interesting avenue of research to explore whether or not there are some machine learning algorithms which could effectively learn the weights of a class of signals given some training data. One could assume that the weights ω are generated from some unknown smooth function f, i.e. $\omega(i) = f(i)$ for $i = 1, \ldots, N$ and apply some nonparametric statistical methods. One could test the quality of the weights by testing to see if weighted ℓ_1 minimization with those learned weights can effectively recover test signals.

Another related problem is to assume that the signals $\boldsymbol{x}$ are being

generated from some parameterized probability distribution $p(\boldsymbol{x})$.[3] While it may make intuitive sense why a weight parameter ω which is monotonically increasing is appropriate for a family of power law decay signals, the manner in which these ω_j components should grow is far from obvious. One may pose the following question: given a signal pdf $p(\boldsymbol{x})$, is there an optimal weight parameter ω? Here, optimal means that with high probability, sparse signals generated from the pdf $p(\boldsymbol{x})$ are recovered from weighted ℓ_1 minimization with weighted parameter ω. If so, how does one compute it? The works [8–10] consider this problem and derive performance guarantees of weighted ℓ_1 minimization for the optimal weight parameter ω. In [8], exact weights were computed for their simpler signal model in which there are two blocks of support and weights ω_1 and ω_2 need to be chosen for each block. In [10] a more general signal model is employed and the authors suggest methods for choosing the weights based on $p(\boldsymbol{x})$. Aside from some relatively simple cases, it is not explicitly known how to compute an optimal set of weights given a model signal pdf $p(\boldsymbol{x})$.

2.7.2 Approximate Projection

The main consequence of the intractability of computing weighted best s-sparse approximations is that we cannot run the IHWT algorithm as each

[3] It should be noted that to optimize the parameters, one typically performs some sort of learning method on training data to optimize the parameters. One common method is to have some training data and use the Expectation Maximization (EM) algorithm to tune the parameters.

iterate requires a projection onto $\mathcal{S}_{\omega,s}$.

To reconcile this issue we compute an approximation to $H_{\omega,s}(\boldsymbol{x})$. Let $\widetilde{H}_{\omega,s}(\boldsymbol{x})$ denote a modified projection operator which sorts the weighted signal $\omega^{-1} \circ \boldsymbol{x}$ [4]. Consult [7] for properties of this weighted thresholding operator.

In what sense is $\widetilde{H}_{\omega,s}$ an approximation to $H_{\omega,s}$? We present the following example to build some intuition. Let $N = 100$ and let ω be given by $\omega(j) = \sqrt{j}$ for $j = 1, \ldots, 100$. Consider the signal $\boldsymbol{x}$ where $x(1) = 10$ and $x(100) = 99$ and equal to 0 otherwise. Then $\omega^{-1} \circ \boldsymbol{x} = [10, 0, \cdots, 0, 9.9]$. Sorting and thresholding we obtain that $\widetilde{\boldsymbol{x}} = \widetilde{H}_{\omega,100}(\boldsymbol{x}) = [10, 0, \cdots, 0]$. Clearly the best weighted 100 sparse approximation is given by $\boldsymbol{x}^* = [0, \cdots, 0, 99]$. In this case, our projection operator $\widetilde{H}_{\omega,s}$ did not compute a very good approximation. However, one can claim that the signal $\boldsymbol{x}$ is a *mis-match* for our weight parameter ω. For signals which "match" the weights more closely, $\widetilde{H}_{\omega,s}$ does a better job of recovering the output of the true projection $H_{\omega,s}$. For example, if $\boldsymbol{x}$ was chosen to be a monotonically decreasing signal, this would match the weight ω and in this case our surrogate $\widetilde{H}_{\omega,s}$ will compute accurate projections.

2.7.3 Experiments

For the remainder of this section, we will be interested in either the approximation or exact recovery of power law distributed signals. To randomly

[4] $A \circ B$ denotes the Hadamard product of A and B.

generate power law signals, we randomly choose integers a, b and formed the power function $f(x) = \frac{a}{x^b}$ and defined our signal $\boldsymbol{x}$ by $\boldsymbol{x}(i) = f(i)$ for $i = 1, \ldots, N$.

We chose our weight parameter ω as follows: the first s-block of coordinates we are relatively certain should be included in our support set as we are dealing with power law signals, thus we set $\omega(1 : s) = 1$. For the second s-block of coordinates we are more uncertain about their inclusion in the signals support set and thus we set $\omega(s + 1 : 2s) = 3$ and we set the tail $\omega(2s + 1 : N) = 10$ for similar reasons. Note that these are still relatively mild weights given the power law prior we have assumed. We further note that given these weights the best (ω, s)-sparse approximation is going to be the actual best s-sparse approximation for s-sparse power law signals.

In the following set of experiments we will test the performance of IHWT for computing (ω, s)-sparse approximations of dense power law decaying signals. For arbitrary dense signals $\boldsymbol{x}$, it requires a combinatorial search to compute the best (ω, s)-sparse approximation. However, for power law decay signals, the best (ω, s)-sparse approximation is simple to compute as it can be performed by choosing the minimal k index such that $\sum_{i=1}^{k} i \leq s$. We note that while the approximate projection operator $\widetilde{H}_{\omega,s}$ will indeed compute the true (ω, s)-sparse approximation of a power law distributed signal, our gradient descent updates $\boldsymbol{x}^{n+1} + \boldsymbol{A}^*(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^n)$ are a priori not going to be power law distributed signals. Therefore in these experiments, we are not only testing the performance of IHWT, but also of this surrogate projection operator $\widetilde{H}_{\omega,s}$.

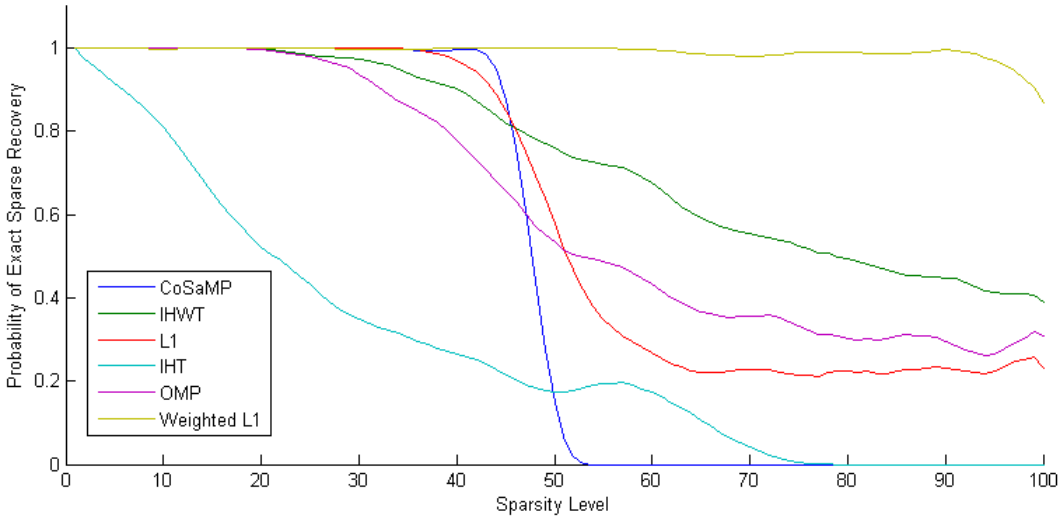

Figure 2.1: Exact Recovery of Randomly Generated Variable s-sparse Power Law Signals using $m = 128$ measurements. Results are averaged over 200 trials. Best viewed in color.

The noisy measurements were $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$ where $\boldsymbol{e}$ is a Gaussian noise vector. To test the quality of our noisy sparse approximation, we computed the normalized error $\frac{\|\boldsymbol{x}_s - \boldsymbol{x}_{\text{approx}}\|_2}{\|\boldsymbol{e}\|_2}$, where $\boldsymbol{x}_s$ is the true best s-sparse approximation and $\boldsymbol{x}_{\text{approx}}$ is the approximation output by our algorithm.

In Figure 2.1 we present the performance of IHWT, CoSaMP [17], IHT [12], OMP [16], ℓ_1 minimization and weighted ℓ_1 minimization for the task of exact sparse recovery using $m = 128$ measurements. In particular we randomly generated $\boldsymbol{A} \in \mathbb{R}^{128\times 256}$ Gaussian sensing matrices, s-sparse power law signals $\boldsymbol{x}_s$ and we have the noise-free measurements $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_s$. We consider a signal to be exactly recovered if the signal approximation and the true underlying signal agree to four decimal places, i.e. $\|\mathrm{x}_{\text{approx}} - x_s\|_2 \leq 10^{-4}$. We averaged over 200 trials. Observe that while IHWT does not exactly recover the sparser power

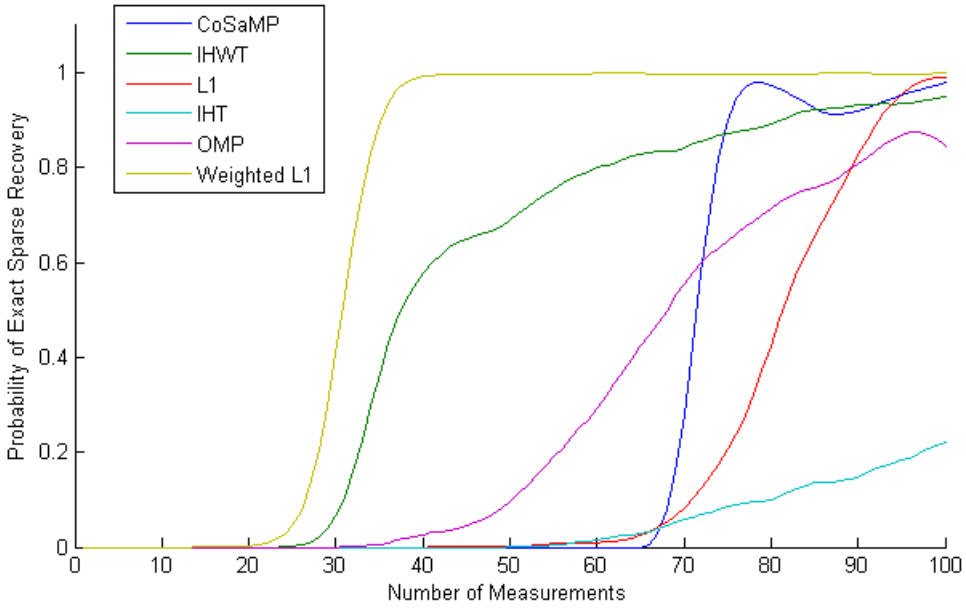

Figure 2.2: Exact Recovery of a fixed sparse $s = 25$ power law distributed signal using a variable number of measurements. Results are averaged over 200 trials. Best viewed in color.

law signals as well as CoSaMP or ℓ_1 minimization, its recovery performance degrades much more gracefully as the sparsity level increases.

In Figure 2.2 we now keep the sparsity level fixed at $s = 25$ and we allow the number of measurements m to vary from 1 to 100. We averaged over 200 runs and we present the probability of exact recovery. Observe the superior performance of IHWT over the other classical greedy sparse approximation algorithms in the undersampling $m = O(s)$ regime.

In the next set of experiments, we tested the noisy sparse recovery performance of IHWT and we compared it again three standard sparse approximation algorithms: CoSaMP, IHT and OMP. We have a fixed number of

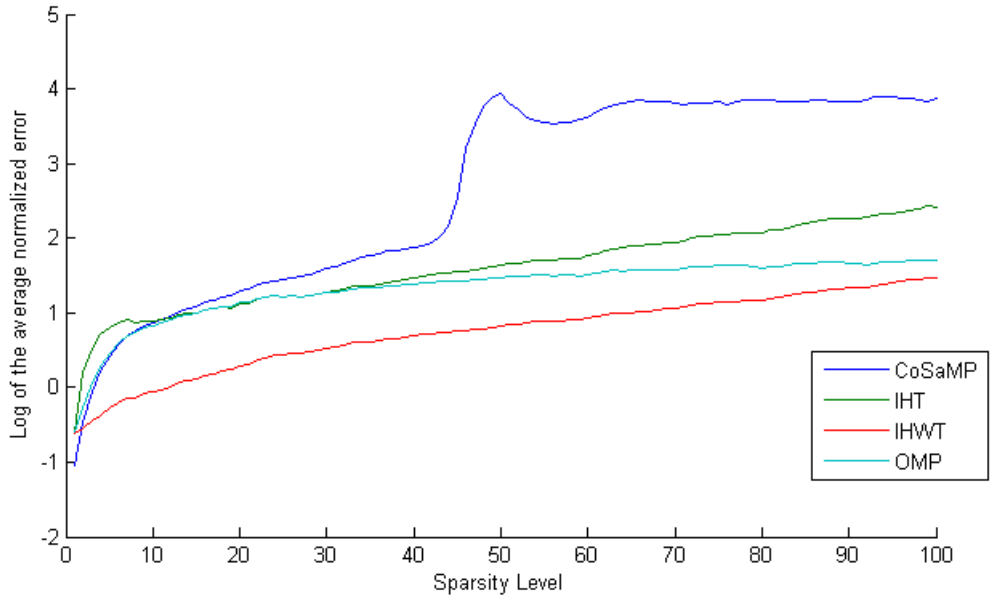

Figure 2.3: The log normalized error averaged over 200 trials of noisy s-sparse approximation of dense Power Law Signals using $m = 128$ measurements. Best viewed in color.

measurements $m = 128$ and we randomly generated $\boldsymbol{A} \in \mathbb{R}^{128\times256}$ Gaussian sensing matrices and we have noisy samples $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$. Note now that $\boldsymbol{x}$ is no longer an s-sparse power law distributed signal but rather it is a *dense* power law distributed signal. In Figure 2.3 we present the log normalized error and log of the standard deviation averaged over 200 trials and in 2.4 we present the log of the standard deviation of the 200 trials. In Figures 2.3 and 2.4 we see the clear performance advantage of IHWT over other greedy algorithms for the task of fixed sparse approximation of power law distributed signals using a fixed number of measurements.

In our final set of experiments, we tested how well we could approximate

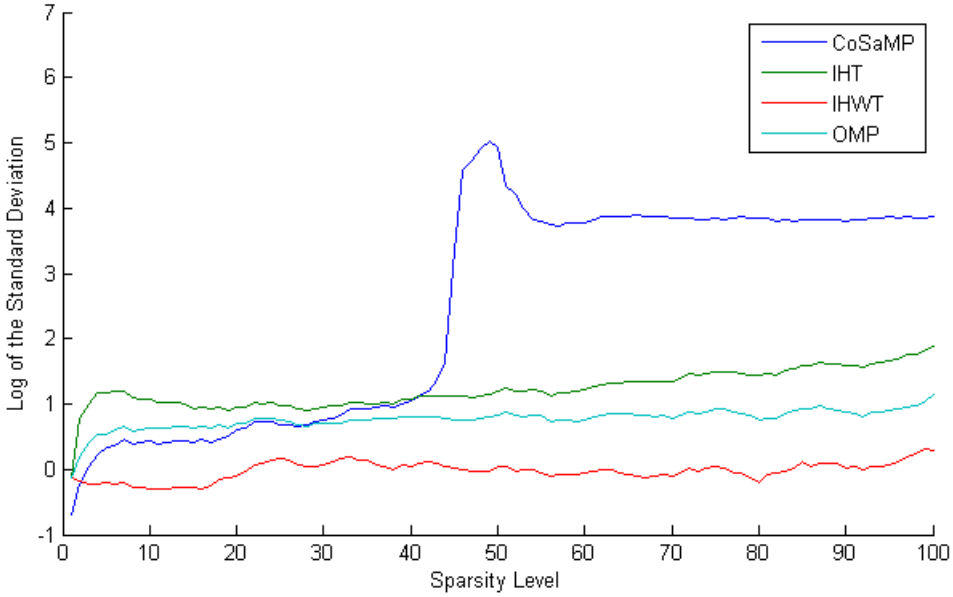

Figure 2.4: The log standard deviation of 200 trials of noisy s-sparse approximation of dense Power Law Signals using $m = 128$ measurements. Best viewed in color.

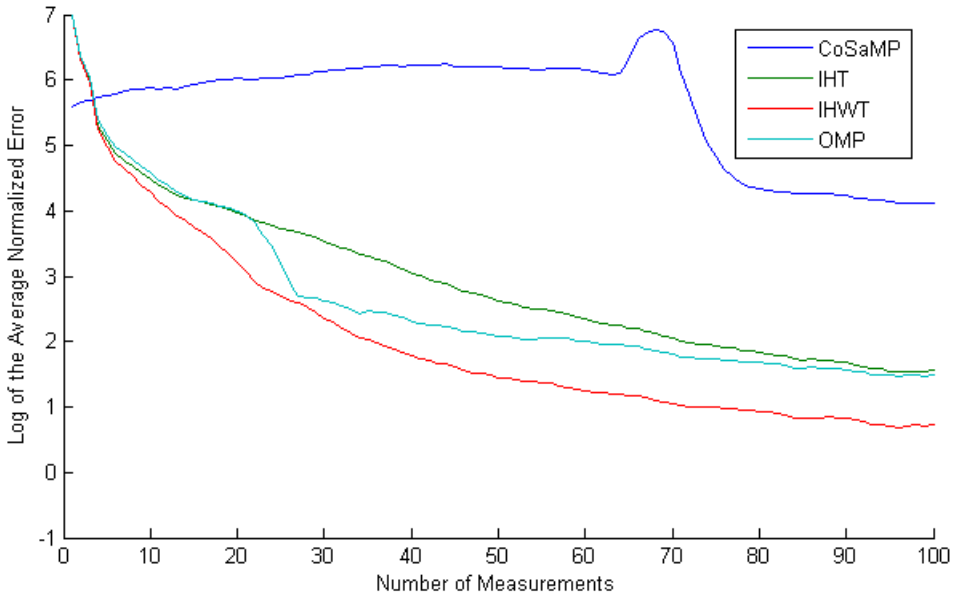

Figure 2.5: The log normalized error averaged over 200 trials of noisy s-sparse approximation of dense Power Law Signals using a variable number of measurements. Best viewed in color.

the best $s = 25$ sparse approximation of a dense power law signal $\boldsymbol{x}$ given a set of noisy measurements $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$ using a variable number of measurements $m = 1, \ldots, 100$. In Figures 2.5 and 2.6 we again see the improved performance of IHWT over other standard greedy sparse approximation algorithms.

2.8 Conclusion and Future Directions

We have presented the IHWT algorithm which is a weighted extension of the IHT algorithm using the weighted sparsity technology developed in [7]. We established theoretical guarantees of IHWT which are weighted analogues of their unweighted counterparts. While not all of the guarantees presented in

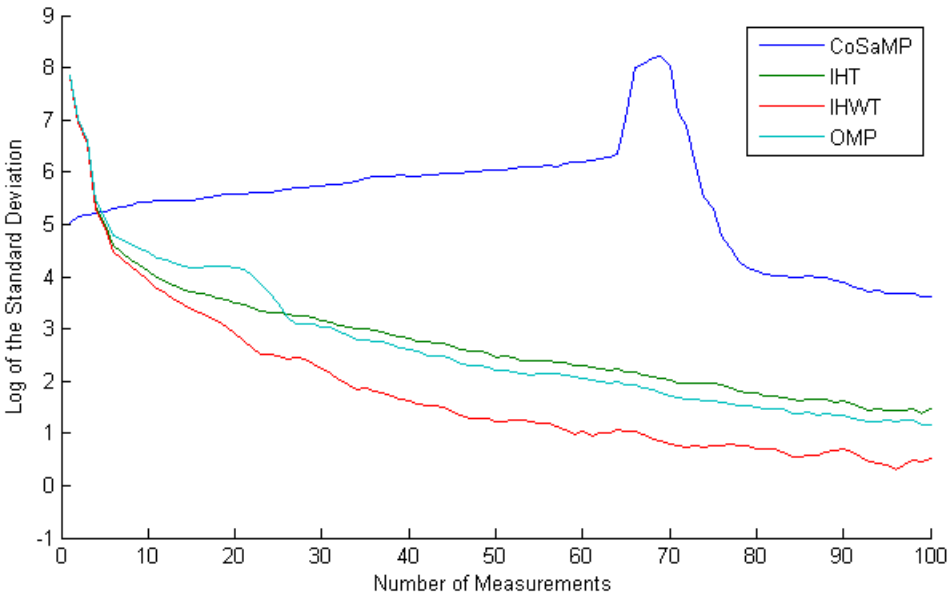

Figure 2.6: The log standard deviation of 200 trials of noisy s-sparse approximation of dense Power Law Signals using a variable number of measurements. Best viewed in color.

[12,19,21] are able to be extended, in certain cases like Prop 2.6.4 and Theorem 2.6.5 we were able to extend the results using the additional hypothesis that the weight parameter ω satisfies $\|\omega\|_\infty \leq O(s)$ for a given weighted sparsity s. This condition allowed us to control the tail $\boldsymbol{x} - \boldsymbol{x}_s$ and instead of obtaining ℓ_p error bounds, we obtained the analogous error bounds in the weighted ℓ_p norms. Empirically to test the performance of IHWT, we implemented a tractable surrogate for the projection onto the space of weighted signals. The numerical experiments also show that the normalized version of IHWT has superior performance to their unnormalized counterparts.

We pose the following open problems:

1. Can the results from [19], which guarantee the convergence of IHT to a local minimizer be extended to IHWT? Here we only extended the guarantee that the IHWT sequence of iterates $(\boldsymbol{x}^n)$ is a contractive sequence.

2. Can we learn the weight parameter ω given some training data $\{\boldsymbol{x}^i\}$ where each $\boldsymbol{x}^i$ is a known signal?

3. Here we simply used the weights from the weighted ℓ_1 minimization problem to be our sparsity weights. However, is there a more optimal choice of weights to reduce the performance gap between IHWT and weighted ℓ_1 minimization for the task of exact sparse recovery?

Chapter 3

Collaborative Filtering: Weighted Matrix Completion

3.1 Overview

Matrix completion has become one of the more active fields in signal processing, enjoying numerous applications to data mining and machine learning tasks. The matrix completion problem is one where we are allowed to observe a small percentage of the entries in a data matrix $\boldsymbol{M}$ and from these known entries, we must infer the values of the remaining entries. This problem is severely ill-posed, particularly so in the high dimensional regime. To this end, one must typically assume some sort of low complexity prior on $\boldsymbol{M}$, i.e. $\boldsymbol{M}$ is a low rank matrix or is well approximated by a low rank matrix. Using this hypothesis a wide range of theoretical guarantees have been established for matrix completion [22–29]. As noted in [3], these articles share a common thread that the recovery guarantees all require that:

- The method of sampling the data matrix $\boldsymbol{M}$ must be done in a uniformly random fashion,
- And that the low-rank matrix $\boldsymbol{M}$ must satisfy a so-called "incoherence" property, which roughly means that the distribution of the entries of the

matrix must have some form of uniform regularity (*thereby allowing the uniform sampling strategy to be effective*).

In [3] it is observed that although the aforementioned articles differ in optimization techniques, ranging from convex relaxation via nuclear norm minimization [23, 24, 29], non-convex alternating minimization [26] and iterative soft thresholding [22], all of these algorithms have exact recovery guarantees using as few as $\Theta(nr \log n)$ observed elements for a square $n \times n$ matrix of rank-r.

One of the central issues in matrix completion is the relationship between the distribution of a matrix's entries and the sampling distribution being employed. For instance, if a matrix is highly incoherent, it has much of its Frobenius norm energy spread throughout its entries in a relatively uniform fashion. To this end, taking a uniformly random sample of this matrix's entries will be a sufficient enough representation to allow for exact recovery. However, if a matrix is highly coherent, in other words, it has much of its Frobenius norm concentrated in a relatively sparse number of its entries, intuitively we understand that a uniform sampling strategy will not yield a sufficiently representative sample to allow for exact recovery.

Up until recently, the exact nature of this relationship between the data matrix $\boldsymbol{M}$ and the sampling distribution $\boldsymbol{p}$ has not been quantified beyond the uniform sampling case. In [3] we see this aforementioned relationship quantified. For the purposes and aims of this chapter, we focus on two particular

results established in [3]:

- If the sampling distribution $\boldsymbol{p}$ is proportional to the sum of the underlying matrix's *leverage scores*, then any arbitrary $n \times n$ rank-r matrix can be recovered from $\Theta(nr \log^2 n)$ observed entries with high probability. The exact recovery guarantee is for the nuclear norm minimization algorithm [30].

- Given a set of weights $\boldsymbol{R}, \boldsymbol{C}$, a sufficiency condition on the sampling distribution $\boldsymbol{p}$ is established. In particular, if the sampling distribution $\boldsymbol{p}$ is proportional to a sum of these $\boldsymbol{R}, \boldsymbol{C}$ weights, then exact recovery guarantees are derived for *weighted nuclear norm minimization* (the particular form of weighted nuclear norm minimization objective was first posed in [31, 32]). Moreover, the benefit of weighted nuclear norm minimization vs. unweighted nuclear norm minimization is quantified with a specific set of weights $\boldsymbol{R}, \boldsymbol{C}$ which are chosen in terms of the sampling distribution $\boldsymbol{p}$.

We are primarily interested in the second result on weighted nuclear norm minimization. We will explore the nature of the relationship between the weights $\boldsymbol{R}, \boldsymbol{C}$ and the empirical sampling distribution $\hat{\boldsymbol{p}}$ as opposed to the true sampling distribution $\boldsymbol{p}$. As previously noted, [3] established the efficacy of weights $\boldsymbol{R}, \boldsymbol{C}$ chosen in a specific fashion in terms of the sampling distribution $\boldsymbol{p}$. **However, we are interested in a setting where the sampling distribution $\boldsymbol{p}$ is not known to us and no prior knowledge**

of $\boldsymbol{p}$ is available and we instead we compute a statistical estimator for $\boldsymbol{p}$. We make the following contributions:

1. We establish a sufficiency condition for the case when the weights $\boldsymbol{R}, \boldsymbol{C}$ are functions of the empirical sampling distribution $\hat{\boldsymbol{p}}$ for the exact recovery of $\boldsymbol{M}$ using weighted nuclear norm minimization.

2. We show that a specific choice of weights $\boldsymbol{R}, \boldsymbol{C}$ as functions of $\hat{\boldsymbol{p}}$ results in a quantifiable relaxation in the exact recovery conditions for weighted nuclear norm minimization vs. unweighted nuclear norm minimization.

3. We numerically demonstrate the healthy robustness of the weighted nuclear norm minimization to the choice of the weights $\boldsymbol{R}, \boldsymbol{C}$, hearkening back to the previous work in non-uniform sampling and weighted matrix completion [31,32]. We also demonstrate the superiority of weighted nuclear norm minimization over unweighted nuclear norm minimization in the non-uniform sampling regime.

To obtain the above two theoretical guarantees we will use a large deviation and a concentration of measure bound from [33] to derive sufficient conditions as to when we may use the empirical sampling distribution $\hat{\boldsymbol{p}}$ as an effective proxy for the true sampling distribution $\boldsymbol{p}$. We use the notation that $a \wedge b := \min(a, b)$ and $a \vee b := \max(a, b)$ throughout.

3.2 Nuclear Norm Minimization and Uniform Sampling

For a given data matrix $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$, let Ω denote a set of observed entries of $\boldsymbol{M}$ where $\Omega \sim \boldsymbol{p}$ and $\boldsymbol{p}$ is a probability mass function on $[n_1] \times [n_2]$. Typically the cardinality of Ω is a small fraction of the $m \cdot n$ total entries. The matrix completion task is the following problem: given the known entries Ω, infer the remaining entries $[n_1] \times [n_2] \setminus \Omega$ of $\boldsymbol{M}$.

The matrix completion task is severely ill-posed and one must regularize/constrain the solution space to obtain a well-posed problem. One popular choice for a low complexity prior for the matrix completion problem is that of low rank: the data matrix $\boldsymbol{M}$ is either low rank or well approximated by a low rank matrix.

One of the main applications of matrix completion is in ratings predictions. In this setting, the rows of our matrix would be the customers and the columns would be their ratings of the various products. Here we may interpret the low rank prior in the following manner: there are a comparatively small (relative to either the number of products or the number of users) set of ratings patterns. Thus from a practical point of view, this low rank prior has some statistical significance which may represent some concentration phenomenon in the observed data.

With this low rank prior in hand, we may pose the following problem, which we refer to as the *rank minimization problem:*

$$\min_{\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}} \operatorname{rank}(\boldsymbol{X}) \text{ subject to } X_{ij} = M_{ij} \text{ for } (i,j) \in \Omega. \tag{3.1}$$

Note that (3.1) is attempting to minimize a non-convex objective over a convex constraint space. Matrix completion is a specific formulation of a more general problem referred to as the *Affine Rank Minimization Problem* (ARMP):

$$\min_{\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}} \operatorname{rank}(\boldsymbol{X}) \text{ subject to } \mathcal{A}(\boldsymbol{X}) = \boldsymbol{y}, \tag{3.2}$$

where $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ is an affine sampling map and $\boldsymbol{y} := \mathcal{A}(\boldsymbol{M})$ is the vector of samples. In the case of matrix completion, the affine map $\mathcal{A} = \mathcal{P}_\Omega$, the projection map onto the Ω indices. The more general ARMP is known to be an NP-Hard problem as ARMP includes sparsest vector (2.10) as a subcase [1, 30].

Using again the convex relaxation heuristic discussed in Chapter 2, one may relax ARMP (3.2) to its tightest convex relaxation. For the case of the rank function, the tightest convex relaxation is the so-called *nuclear norm*:

Definition 3.2.1. Let a matrix $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ have singular value decomposition given by $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$. Then the nuclear norm is defined as:

$$\|\boldsymbol{M}\|_* := \sum_{i=1}^{\operatorname{rank}(\boldsymbol{M})} \Sigma_{ii}, \tag{3.3}$$

i.e. the sum of the singular values of $\boldsymbol{M}$.

Intuitively, one may view the nuclear norm as the tightest convex relaxation of the rank function in the following manner: the rank of $\boldsymbol{M}$ is merely the ℓ_0 pseudo-norm of the vector of diagonal entries of $\boldsymbol{\Sigma}$ from the SVD of $\boldsymbol{M}$, and we know that the ℓ_1 norm is the tightest convex relaxation of the ℓ_0

norm, and the ℓ_1 norm of the vector of singular values will be the sum of the singular values of $\boldsymbol{M}$, precisely the nuclear norm. For a proof consult [30].

In [30] the following guarantee was established for (3.2) (which we paraphrase using our notation):

Theorem 3.2.1. *Suppose that $\delta_{2r} < 1$ for some integer $r \geq 1$. Then $\boldsymbol{M}$ is the only matrix of rank at most r satisfying $\mathcal{A}(\boldsymbol{M}) = \boldsymbol{y}$.*

Where the δ_{2r} is the RIP constant of order $2r$ for $\mathcal{A}$. Thus [30] established that (3.2) is well-posed subject to the sampling operator $\mathcal{A}$ acting as an approximate isometry, mirroring much of the theoretical developments in compressive sensing.

Relaxing (3.1) yields the nuclear norm minimization based matrix completion problem:

$$\min_{\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}} \|\boldsymbol{X}\|_* \text{ subject to } X_{ij} = M_{ij} \text{ for } (i,j) \in \Omega, \tag{3.4}$$

We may view (3.4) as a specific case of the nuclear norm relaxed ARMP where the affine sampling operator $\mathcal{A} = \mathcal{P}_\Omega$. Note that $\mathcal{P}_\Omega$ is in general not going to be an isometry for any sampling rate m which allows for a sufficient compression of the original data matrix $\boldsymbol{M}$. To this end, the guarantee presented in Theorem (3.2.1) is not applicable to matrix completion.

However, with the theoretical properties of the nuclear norm relaxation heuristic firmly established, a flood of matrix completion results followed, starting with [23]. Soon after, [24] established the following guarantee for (3.4):

Theorem 3.2.2. *Let $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ be a fixed matrix of rank $r = O(1)$ with incoherence parameter μ_0. Write $n := \min(n_1, n_2)$. Suppose we observe m entries of $\boldsymbol{M}$ with locations sampled uniformly at random. Then there is a positive numerical constant C such that if:*

$$m \geq C\mu_0^4 n(\log n)^2, \tag{3.5}$$

then $\boldsymbol{M}$ is the unique solution to (3.4) with probability at least $1 - n^{-3}$. In other words: with high probability, nuclear norm minimization recovers all the entries of $\boldsymbol{M}$ with no error.

And for general rank r-matrices, Theorem 3.2.2 can be extended to yield exact recovery with high probability for the following sampling lower bound:

$$m \geq C\mu_0^4 nr^2(\log n)^2. \tag{3.6}$$

The *incoherence property* for a matrix $\boldsymbol{M}$ is defined as:

Definition 3.2.2. Let $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ be a rank-r matrix with SVD given by $\boldsymbol{M} = U\Sigma V^T$. Then $\boldsymbol{M}$ is said to satisfy the *incoherence property* with *incoherence parameter* $\mu_0 > 0$ if:

$$\|\mathcal{P}_U(\boldsymbol{e}_a)\|^2 \leq \frac{\mu_0 r}{n_1}, \quad \|\mathcal{P}_V(\boldsymbol{e}_b)\|^2 \leq \frac{\mu_0 r}{n_2}, \tag{3.7}$$

for all $a \in [n_1], b \in [n_2]$ where $\boldsymbol{e}_a, \boldsymbol{e}_b$ represent the respective standard Euclidean basis vectors.

Note that a lower incoherence parameter implies that the matrix $\boldsymbol{M}$ has column and row spaces which are not highly correlated with the respective standard Euclidean basis vectors. For example, if the column space of $\boldsymbol{M}$ is spanned by vectors whose entries all have magnitude $1/\sqrt{n_1}$, then $\frac{n_1}{r}\|\mathcal{P}_U(\boldsymbol{e}_a)\|^2$ attains the minimal value 1. In other words: *when there is a uniform distribution of energy, the incoherence is minimized.* On the other hand, if the column space of $\boldsymbol{M}$ contains a standard basis vector $\boldsymbol{e}_a$ for some $a \in [n_1]$, then $\frac{n_1}{r}\|\mathcal{P}_U(\boldsymbol{e}_a)\|^2$ attains the maximal value $\frac{n_1}{r}$. In other words, $\boldsymbol{M}$ having a lower coherence parameter translates to $\boldsymbol{M}$ not having any pathological distribution of its energy, for example $\boldsymbol{M}$ being a zero matrix except for a single entry. A lower incoherence parameter guarantees some sort of regularity to the distribution of the entries of $\boldsymbol{M}$, thus enabling a vanilla sampling strategy such as uniform sampling to allow for exact recovery. Thus for the first generation of matrix completion algorithms, a low coherence prior was typically assumed in conjunction with the low rank prior in order to guarantee that the matrix $\boldsymbol{M}$ would not be in the null space of the sampling operator $\mathcal{P}_\Omega$.

Related to [24], [29] provided a simplified and more streamlined analysis of the theoretical guarantees for (3.4). Other techniques have been developed to solve (3.4) using a range of different optimization techniques, from non-convex alternating minimization [26] to iterative soft thresholding [22]. These methods all enjoy exact recovery guarantees for a square $n \times n$ rank-r matrix in the uniform sampling regime using as few as $\Theta(nr \log n)$ known samples. Despite their differences in optimization techniques, all of these results share

the common hypotheses of *uniform sampling* and *incoherence*. Indeed, the incoherence hypothesis is strongly related to the uniform sampling hypothesis. If $\boldsymbol{M}$ has a small coherence parameter $\mu_0(\boldsymbol{M})$, then as noted above, the distribution of the entries of $\boldsymbol{M}$ is sufficiently regular in the sense that $\boldsymbol{M}$ does not have a pathological concentration of energy in a small number of entries. Not having a pathological concentration of energy in a small number of entries enables a uniform sampling to be sufficient to achieve exact recovery.

3.3 Non-uniform Matrix Completion

Early results on matrix completion [22–26, 29] have illuminated a connection between the nature of uniform sampling and the coherence of a matrix $\boldsymbol{M}$. In particular, the aforementioned articles illuminate the quantitative relationship between uniform sampling and a matrix's incoherence parameter, and when exact recovery using (3.4) is possible. Up until recently, the exact nature of this relationship between $\boldsymbol{M}$'s statistics and the sampling distribution $\boldsymbol{p}$ has not been quantified beyond the uniform sampling case. In [3] a general relationship between a specific set of statistics of the matrix $\boldsymbol{M}$, referred to as *leverage scores* and any arbitrary sampling distribution $\boldsymbol{p}$ has been established. From [3] recall that the *leverage scores* of a matrix $\boldsymbol{M}$ are defined as:

Definition 3.3.1. (Leverage Scores) For an $n_1 \times n_2$ real valued matrix $\boldsymbol{M}$ of rank-r whose rank-r SVD is given by $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, its (normalized) leverage

scores–$\mu_i(\boldsymbol{M})$ for any row i, and $\nu_j(\boldsymbol{M})$ for any column j–are defined as:

$$\mu_i(\boldsymbol{M}) := \frac{n_1}{r}\|\boldsymbol{U}^T\boldsymbol{e}_i\|_2^2, i = 1, \ldots, n_1,$$
$$\nu_j(\boldsymbol{M}) := \frac{n_2}{r}\|\boldsymbol{V}^T\boldsymbol{e}_j\|_2^2, j = 1, \ldots, n_2,$$

where $\boldsymbol{e}_i$ denotes the i-th standard basis vector with appropriate dimension. The coherence of a matrix, hereby denoted as $\mu_0(\boldsymbol{M})$ serves as a uniform upper bound for the leverage scores:

$$\mu_0(\boldsymbol{M}) \geq \max_{i,j}\{\mu_i(\boldsymbol{M}), \nu_j(\boldsymbol{M})\}.$$

Then the following theorem was established:

Theorem 3.3.1. *(Theorem 2 of [3]) Let $\boldsymbol{M} = (M_{ij})$ be an $n_1 \times n_2$ matrix of rank-r, and suppose that its elements M_{ij} are observed only over a subset of elements $\Omega \subset [n_1] \times [n_2]$. There is a universal constant $c_0 > 0$ such that, if each element (i, j) is independently observed with probability p_{ij}, and p_{ij} satisfies:*

$$p_{ij} \geq \min\left(c_0 \frac{(\mu_i(\boldsymbol{M}) + \nu_j(\boldsymbol{M}))r\log^2(n_1 + n_2)}{\min(n_1, n_2)}, 1\right), \tag{3.8}$$
$$p_{ij} \geq \frac{1}{\min(n_1, n_2)^{10}} \tag{3.9}$$

then $\boldsymbol{M}$ is the unique optimal solution to the nuclear norm minimization problem (3.4) with probability at least $1 - 5(n_1 + n_2)^{-10}$.

Note: The expected number of samples will be $O(\max(n_1, n_2)r\log^2(n_1 + n_2))$.

Observe that Theorem 3.3.1 establishes a quantitative relationship between a matrix's leverage scores and the sampling distribution $\boldsymbol{p}$ which is sufficient to allow for exact recovery using (3.4).

Consider now a different scenario, one in which the sampling distribution $\boldsymbol{p}$ and the underlying matrix's leverage scores $\{\mu_i(\boldsymbol{M})\}_{i=1}^{n_1}, \{\nu_j(\boldsymbol{M})\}_{j=1}^{n_2}$ do not align according to (3.8). One technique to remedy this situation is to design a transformation $\boldsymbol{M} \mapsto \bar{\boldsymbol{M}}$ so that *we may adjust the leverage scores to align with the sampling distribution* $\boldsymbol{p}$. We choose weights of the form $\boldsymbol{R} := \text{diag}(R_1, \ldots, R_{n_1}) \in \mathbb{R}^{n_1 \times n_1}, \boldsymbol{C} := \text{diag}(C_1, \ldots, C_{n_2}) \in \mathbb{R}^{n_2 \times n_2}$. Using these parameterized weights, we will use $\boldsymbol{M} \mapsto \boldsymbol{RMC}$ as our transformation which will adjust leverage scores of $\boldsymbol{M}$. In [31] a weighted nuclear norm objective was proposed. Following [3, 32], we will be considering the following *weighted nuclear norm minimization problem*:

$$\bar{\boldsymbol{M}} = \underset{\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}}{\text{argmin}} \|\boldsymbol{RXC}\|_* \text{ subject to } X_{ij} = M_{ij}, \text{ for } (i,j) \in \Omega. \tag{3.10}$$

Let $\boldsymbol{USV}^T$ denote the rank-r SVD of $\boldsymbol{M}$. In [3] the row leverage scores of the transformed matrix $\bar{\boldsymbol{M}} = \boldsymbol{RMC}$ were upper bounded by the weights $\boldsymbol{R}, \boldsymbol{C}$, the coherence $\mu_0(\boldsymbol{M})$ and the singular values of $\boldsymbol{RU}$ (and respectively $\boldsymbol{V^TC}$; the analysis is identical). To bound the singular values of $\boldsymbol{RU}$ a linear program is analyzed. It is known that the extrema of linear programs are obtained at the boundary of the feasible set. In [3], it was established that given any set of weights $\boldsymbol{R}, \boldsymbol{C}$, the corresponding support sets $\mathcal{S}_r, \mathcal{S}_c$ are the $\lfloor n_1/(\mu_0 r) \rfloor, \lfloor n_2/(\mu_0 r) \rfloor$ entries of least magnitude of $\boldsymbol{R}, \boldsymbol{C}$, respectively. We let

$\mathcal{S}_r, \mathcal{S}_c$ denote the corresponding indices of the $\lfloor n_1/(\mu_0 r)\rfloor, \lfloor n_2/(\mu_0 r)\rfloor$ entries of least magnitude of $\boldsymbol{R}, \boldsymbol{C}$, respectively.

In [3], a quantitative relationship between the weights $\boldsymbol{R}, \boldsymbol{C}$ and the sampling distribution $\boldsymbol{p}$ which guarantees that (3.10) is an optimal solution was established[1]:

Theorem 3.3.2. *(Theorem 7 in [3]) Let $\boldsymbol{M} = (M_{ij})$ be an $n \times n$ matrix of rank-r, and suppose that its elements M_{ij} are observed only over a subset of elements $\Omega \subset [n]\times[n]$. Without loss of generality, assume $R_1 \leq R_2 \leq \cdots \leq R_n$ and $C_1 \leq C_2 \leq \cdots \leq C_n$. There exists a universal constant c_0 such that $\boldsymbol{M}$ is the unique optimum to* (3.10) *with probability at least $1 - 5(2n)^{-10}$ provided that for all $(i,j) \in [n] \times [n], p_{ij} \geq n^{-10}$ and:*

$$p_{ij} \geq c_0 \left(\frac{R_i^2}{\sum_{i'=1}^{\lfloor (n/(\mu_0 r)\rfloor} R_{i'}^2} + \frac{C_j^2}{\sum_{j'=1}^{\lfloor (n/(\mu_0 r)\rfloor} C_{j'}^2} \right) \log^2(2n). \tag{3.11}$$

Note that for monotonically increasing weights $\boldsymbol{R}, \boldsymbol{C}$ the corresponding support sets $\mathcal{S}_r, \mathcal{S}_c$ are merely the first $\lfloor n/(\mu_0 r)\rfloor$ indices, respectively.

3.4 Main Results

In what follows we shall assume that our sampling distribution $\boldsymbol{p}$ has a product form $p_{ij} = p_i^r p_j^c$ for all $(i,j) \in [n_1] \times [n_2]$. Furthermore, we will consider the following *two-stage sampling model*:

[1] We state Theorem 3.3.2 in the square $n \times n$ case for our purposes.

- Stage 1 (Empirical Sampling Distribution): We sample the distribution $\boldsymbol{p}$ with m times independently with replacement, but the corresponding entries of the data matrix $\boldsymbol{M}$ are not revealed to us. In other words, we are *sampling the sampling distribution*, but not the underlying matrix $\boldsymbol{M}$.

- Stage 2 (Sampling the Matrix): We then, independent of the first stage, sample the matrix $\boldsymbol{M}$ using the independent Bernoulli model for each entry $(i, j) \in [n_1] \times [n_2]$.

Note that this two stage sampling models allows one to sample the sampling distribution $\boldsymbol{p}$ without revealing the entries of $\boldsymbol{M}$. In this manner we may design weights $\boldsymbol{R}, \boldsymbol{C}$ which depend on the empirical sampling distribution $\hat{\boldsymbol{p}}$ and obtain matrix completion guarantees for these weights in the usual (stage two) independent Bernoulli sampling model that has been typically used in the matrix completion literature.

We present stage one sampling bounds which will allow $\hat{\boldsymbol{p}}$ to be used as an empirical proxy for $\boldsymbol{p}$ to design weights $\boldsymbol{R}, \boldsymbol{C}$ for (3.10) and obtain exact recovery with high probability. To this end, we establish the following two empirical estimation lemmas, which will serve as the foundation to our matrix completion guarantees. The first is a *one sided large deviation bound*:

Lemma 3.4.1. *Let $\boldsymbol{p}$ denote a probability mass function on $[n_1] \times [n_2]$ and suppose $\boldsymbol{p}$ has a product form, i.e. for all $(i, j) \in [n_1] \times [n_2] : p_{ij} = p_i^r p_j^c$ for $\boldsymbol{p^r}, \boldsymbol{p^c}$ probability mass functions on $[n_1], [n_2]$, respectively. Let $X_1, \ldots, X_m \overset{i.i.d}{\sim} \boldsymbol{p}$ be*

a sequence of m i.i.d samples. For any $\alpha \in (0, (\min_{i\in[n_1]} p_i^r \vee \min_{j\in[n_2]} p_j^c)^{-1})$ and $\epsilon \in (0,1)$, if the number of samples m is chosen such that:

$$m = \frac{1}{2}\left(\alpha \min_{i\in[n_1]} p_i^r \wedge \min_{j\in[n_2]} p_j^c\right)^{-2} \log(\epsilon^{-1}(n_1+n_2)), \tag{3.12}$$

then with probability at least $1-\epsilon$ we have that for all $(i,j) \in [n_1]\times[n_2]$:

$$p_{ij} \geq \frac{1}{(1+\alpha)^2}\hat{p}_{ij}. \tag{3.13}$$

We also establish the following *two sided empirical bound* for the estimation of product distributions:

Lemma 3.4.2. *Let $\boldsymbol{p}$ denote a probability mass function on $[n_1]\times[n_2]$ and suppose $\boldsymbol{p}$ has a product form, i.e. for all $(i,j) \in [n_1]\times[n_2] : p_{ij} = p_i^r p_j^c$ for $\boldsymbol{p^r}, \boldsymbol{p^c}$ probability mass functions on $[n_1], [n_2]$, respectively. Let $X_1, \ldots, X_m \overset{i.i.d}{\sim} \boldsymbol{p}$ be a sequence of m i.i.d samples. For any $\alpha \in (0, (\min_{i\in[n_1]} p_i^r \vee \min_{j\in[n_2]} p_j^c)^{-1})$ and $\epsilon \in (0,1)$, if the number of samples m is chosen such that:*

$$m = \frac{1}{2}\left(\alpha \min_{i\in[n_1]} p_i^r \wedge \min_{j\in[n_2]} p_j^c\right)^{-2} \log(2\epsilon^{-1}(n_1+n_2)), \tag{3.14}$$

then with probability at least $1-\epsilon$ we have that for all $(i,j) \in [n_1]\times[n_2]$:

$$\frac{1}{(1+\alpha)^2}\hat{p}_{ij} \leq p_{ij} \leq \frac{1}{(1-\alpha)^2}\hat{p}_{ij}. \tag{3.15}$$

Note that Lemmas 3.4.1 and 3.4.2 are general results for the empirical estimation of any distribution $\boldsymbol{p}$ over $[n_1]\times[n_2]$ which has a product form. Recall that the sampling model employed in [3] is a sequence of $n_1 \cdot n_2$ independent Bernoulli random variables, with each Bernoulli random variable

having success probability p_{ij} for $(i,j) \in [n_1] \times [n_2]$. Therefore, $\boldsymbol{p}$ may not be a probability matrix on $[n_1] \times [n_2]$ as it may not sum to 1. To this end, we note that when we sample $\boldsymbol{p}$, we are really sampling the normalized matrix $\frac{1}{\sum_{i,j} p_{ij}} \boldsymbol{p}$. So our empirical estimator $\hat{\boldsymbol{p}}$ is estimating the normalized probability matrix $\frac{1}{\sum_{i,j} p_{ij}} \boldsymbol{p}$ and not $\boldsymbol{p}$ itself. Therefore, in order to apply the above lemmas we must account for this normalization constant.

Using the above, we will obtain two weighted matrix completion guarantees. For simplicity, we will prove all our results for the case when $\boldsymbol{M}$ is a square $n \times n$ matrix. The first guarantee will be a sufficiency condition for the weights $\boldsymbol{R}, \boldsymbol{C}$ in terms of the empirical estimator $\hat{\boldsymbol{p}}$ which will ensure exact recovery by weighted nuclear norm minimization with high probability:

Theorem 3.4.3. *Let $\boldsymbol{M} = (M_{ij})$ be an $n \times n$ matrix of rank-r, and suppose that its elements M_{ij} are observed only over a subset of elements $\Omega \subset [n] \times [n]$, Let $\epsilon \in (0,1)$ be arbitrary. Suppose that there exists*

$$\alpha \in (0, (\min_{i \in [n]} p_i^r / (\sum_{i \in [n]} p_i^r) \vee \min_{j \in [n]} p_j^c / (\sum_{j \in [n]} p_j^c))^{-1})$$

and some universal constant c_0 such that for all indices $(i,j) \in [n] \times [n]$ the weights $\boldsymbol{R}, \boldsymbol{C}$ satisfy the following inequalities:

$$\hat{p}_{ij} \geq \frac{(1+\alpha)^2}{\sum_{ij} p_{ij}} c_0 \left(\frac{R_i^2}{\sum_{i' \in \mathcal{S}_r} R_{i'}^2} + \frac{C_j^2}{\sum_{j' \in \mathcal{S}_c} C_{j'}^2} \right) \log^2(2n), \tag{3.16}$$

where $\mathcal{S}_r, \mathcal{S}_c$ denote the $\lfloor n/(\mu_0 r) \rfloor$ entries of least magnitude of $\boldsymbol{R}, \boldsymbol{C}$, respectively. If the number of stage one samples m is chosen such that:

$$m = \frac{1}{2} \left(\alpha \min_{i \in [n]} \frac{p_i^r}{\sum_{i=1}^n p_i^r} \wedge \min_{j \in [n]} \frac{p_j^c}{\sum_{j=1}^n p_j^c} \right)^{-2} \log(2\epsilon^{-1} n)$$

and if for all $(i,j) \in [n] \times [n], p_{ij} \geq n^{-10}$*, then with probability at least* $(1 - 5(2n)^{-10})(1-\epsilon)$*,* $\boldsymbol{M}$ *is unique optimum to* (3.10)*, where* Ω *is obtained via the usual (stage two) independent, entry-wise Bernoulli sampling of* $\boldsymbol{M}$*.*

Our second weighted matrix completion guarantee will be for the exact recovery properties of a set weights $\boldsymbol{R}, \boldsymbol{C}$ explicitly defined in terms of the empirical distribution $\hat{\boldsymbol{p}}$:

Theorem 3.4.4. *Let* $\boldsymbol{M}$ *be a square* $n \times n$ *rank-*r *matrix with coherence* μ_0*. Consider the weights defined by:*

$$R_i = \sqrt{\frac{1}{n}\hat{p}_i^r \sum_{j' \in \mathcal{S}_c} \hat{p}_{j'}^c} \text{ for } i = 1, \ldots, n, \tag{3.17}$$

$$C_j = \sqrt{\frac{1}{n}\hat{p}_j^c \sum_{i' \in \mathcal{S}_r} \hat{p}_{i'}^r} \text{ for } j = 1, \ldots, n, \tag{3.18}$$

where $\mathcal{S}_r, \mathcal{S}_c$ *denote the* $\lfloor n/(\mu_0 r) \rfloor$ *entries of* $\hat{\boldsymbol{p}}^{\boldsymbol{r}}, \hat{\boldsymbol{p}}^{\boldsymbol{c}}$ *of least magnitude, respectively. Suppose that there exists*

$$\alpha \in (0, (\min_{i \in [n]} p_i^r / (\sum_{i \in [n]} p_i^r) \vee \min_{j \in [n]} p_j^c / (\sum_{j \in [n]} p_j^c))^{-1})$$

such that the (unnormalized) matrix $\boldsymbol{p}$ *satisfies for all* $(i,j) \in [n] \times [n]$ *and the sets* $\mathcal{S}_r^*, \mathcal{S}_c^*$ *which denote the* $\lfloor n/(\mu_0 r) \rfloor$ *entries of* $\boldsymbol{p}^{\boldsymbol{r}}, \boldsymbol{p}^{\boldsymbol{c}}$ *of least magnitude, respectively satisfies the following:*

$$p_j^c \sum_{i' \in \mathcal{S}_r^*} p_{i'}^r \geq c_0 \frac{2(1+\alpha)^2}{(1-\alpha)^2} \log^2(2n), \tag{3.19}$$

$$p_i^r \sum_{j' \in \mathcal{S}_c^*} p_{j'}^c \geq c_0 \frac{2(1+\alpha)^2}{(1-\alpha)^2} \log^2(2n). \tag{3.20}$$

If the number of stage one samples m is chosen such that:

$$m = \frac{1}{2}\left(\alpha \min_{i\in[n]} \frac{p_i^r}{\sum_{i=1}^n p_i^r} \wedge \min_{j\in[n]} \frac{p_j^c}{\sum_{j=1}^n p_j^c}\right)^{-2} \log(4\epsilon^{-1}n),$$

then with probability at least $(1-5(2n)^{-10})(1-\epsilon)$, $\boldsymbol{M}$ *is unique optimum to* (3.10)*, where* Ω *is obtained via the usual (stage two) independent, entry-wise Bernoulli sampling of* $\boldsymbol{M}$*.*

Note: Unweighted nuclear norm minimization attains exact recovery under the condition that for all $(i,j) \in [n] \times [n]$:

$$p_i^r p_j^c \gtrsim \frac{\mu_0 r}{n} \log^2(2n). \tag{3.21}$$

However as Theorem 3.4.4 establishes, weighted nuclear norm minimization with choice of weights (3.17) and (3.18) attains exact recovery subject to the less restrictive sufficient recovery condition that:

$$p_j^c \sum_{i'\in\mathcal{S}_r^*} p_{i'}^r \gtrsim \log^2(2n),$$
$$p_i^r \sum_{j'\in\mathcal{S}_c^*} p_{j'}^c \gtrsim \log^2(2n).$$

This is precisely the condition from [3].

3.5 Empirical Estimation

We consider probability mass functions $\boldsymbol{p}$ on $[n_1] \times [n_2]$ which have a product form $p_{ij} = p_i^r p_j^c$ for $(i,j) \in [n_1] \times [n_2]$. We will sample this distribution with replacement m times. The $X_1, \ldots, X_m \overset{i.i.d}{\sim} \boldsymbol{p}$ samples are row and column

pairs, i.e. $X_k \in [n_1] \times [n_2]$ for each $k = 1, \ldots, m$. We may define the *row and column empirical estimators*:

Definition 3.5.1. The row and column empirical estimators $\hat{\boldsymbol{p}}^{\boldsymbol{r}}, \hat{\boldsymbol{p}}^{\boldsymbol{c}}$, respectively are defined as:

$$\hat{p}_i^r := \frac{1}{m} \sum_{k=1}^{m} \delta_i^r(X_k), \text{ for } i \in [n_1], \tag{3.22}$$

$$\hat{p}_j^c := \frac{1}{m} \sum_{k=1}^{m} \delta_j^c(X_k), \text{ for } j \in [n_2], \tag{3.23}$$

where for any X_k:

$$\delta_i^r(X_k) = \begin{cases} 1 & \text{if } X_k \text{ is from row } i, \\ 0 & \text{otherwise.} \end{cases}$$

$$\delta_j^c(X_k) = \begin{cases} 1 & \text{if } X_k \text{ is from column } j, \\ 0 & \text{otherwise.} \end{cases}$$

For the remainder we will allow $\hat{\boldsymbol{p}}$ denote the empirical product estimate, i.e. $\hat{\boldsymbol{p}} = \hat{\boldsymbol{p}}^{\boldsymbol{r}} \hat{\boldsymbol{p}}^{\boldsymbol{c}}$.

Observe that in (3.22) and (3.23) each component of our row and column empirical estimators involve a sum of independent, bounded in $[0, 1]$ random variables as $\delta_i^r(X_k), \delta_j^c(X_k) \in \{0, 1\}$ for any $(i, j, k) \in [n_1] \times [n_2] \times [m]$. In this situation, we may use *Hoeffding's inequalities* [33] to obtain some probabilistic approximation guarantees. For our purposes, we will be using two forms of Hoeffding's inequalities: a one sided large deviation bound and a two sided concentration of measure bound.

Theorem 3.5.1. *(Hoeffding Inequalities) Let $Z_1, \ldots, Z_m$ be independent random variables such that each $Z_i \in [a_i, b_i]$ with probability 1. Let $S_m = \sum_{i=1}^{m} Z_i$. Then for any $t > 0$ we have:*

$$\Pr[S_m - \mathbb{E}[S_m] \geq t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right), \tag{3.24}$$

$$\Pr[|S_m - \mathbb{E}[S_m]| \geq t] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right). \tag{3.25}$$

For any $i \in [n_1]$, we may define m random variables $Z_{i,k} := \delta_i^r(X_k)$ for $k = 1, \ldots, m$. Note that each random variable $Z_{i,k}$ only takes values in $\{0, 1\}$ and thus is bounded in $[0, 1]$ with probability 1. As each X_k is merely a row and column index, and each δ_i^r, δ_j^c are row and column indicator functions, we have that any set of the $Z_{i,k}$'s (and similarly for the column case) is an independent set of random variables. Therefore the hypotheses of Theorem 3.5.1 are satisfied. For each $i \in [n_1]$ we may define the sum $S_{i,m}^r := \sum_{k=1}^{m} Z_{i,k}$. Each $S_{i,m}^r$ has expected value $\mathbb{E}[S_{i,m}^r] = mp_i^r$. Analogous results hold for the column case. With the above pair of Hoeffding inequalities in hand, we are now ready to establish our main lemmas. For the proof of Lemma 3.4.1 we will apply (3.24) and for the proof of Lemma 3.4.2 we will apply (3.25).

3.5.1 Proof Lemma 3.4.1

Proof. We start our proof by analyzing empirical estimation of the row distribution; the analysis for the column distribution will be identical. For any $i \in [n_1], \alpha > 0$, choosing $t = \alpha \min_{i \in [n_1]} p_i^r$, by (3.24) we have that:

$$\Pr[\hat{p}_i^r - p_i^r \geq \alpha \min_{i \in [n_1]} p_i^r] \leq \exp(-2(\alpha \min_{i \in [n_1]} p_i^r)^2 m). \tag{3.26}$$

We repeat the analysis for the column case: choose $t = \alpha \min_{j\in[n_2]} p_j^c$, then analogously

$$\Pr[\hat{p}_j^c - p_j^c \geq \alpha \min_{j\in[n_2]} p_j^c] \leq \exp(-2(\alpha \min_{j\in[n_2]} p_j^c)^2 m). \tag{3.27}$$

For any $i \in [n_1]$ let E_i^r denote the event that $\hat{p}_i^r - p_i^r \geq \alpha \min_{i\in[n_1]} p_i^r$ and for any $j \in [n_2]$ let E_j^c denote the event that $\hat{p}_j^c - p_j^c \geq \alpha \min_{j\in[n_2]} p_j^c$.

We must choose $\alpha > 0$ such that the bounds in (3.26), (3.27) are nontrivial. In particular, any two probability vectors cannot have their components differ by more than 1. Therefore, we require that α satisfies:

$$\alpha \min_{i\in[n_1]} p_i^r \leq 1 \text{ and } \alpha \min_{j\in[n_2]} p_j^c \leq 1.$$

To this end it suffices to choose $\alpha \in (0, (\min_{i\in[n_1]} p_i^r \vee \min_{j\in[n_2]} p_j^c)^{-1})$.

By (3.26), (3.27) and the Union Bound we have that:

$$\begin{aligned}
&\Pr\left[\text{For some } (i,j) \text{ the event } E_i^r \text{ or } E_j^c \text{ occurs}\right] &(3.28)\\
&\leq \left(n_1 \exp(-2(\alpha \min_{i\in[n_1]} p_i^r)^2 m) + n_2 \exp(-2(\alpha \min_{j\in[n_2]} p_j^c)^2 m\right)\\
&\leq (n_1+n_2)\exp(-2(\alpha \min_{i\in[n_1]} p_i^r \wedge \min_{j\in[n_2]} p_j^c)^2 m). &(3.29)
\end{aligned}$$

Observe that (3.29) immediately yields that with probability at least $1-(n_1+n_2)\exp(-2(\alpha \min_{i\in[n_1]} p_i^r \wedge \min_{j\in[n_2]} p_j^c)^2 m)$ for any $(i,j) \in [n_1]\times[n_2]$ we have that the following bounds hold:

$$\hat{p}_i^r - p_i^r \leq \alpha \min_{i\in[n_1]} p_i^r, \tag{3.30}$$

$$\hat{p}_j^c - p_j^c \leq \alpha \min_{j\in[n_2]} p_j^c. \tag{3.31}$$

Therefore with probability at least $1-(n_1+n_2)\exp(-2(\alpha \min p_{i,j})^2 m)$ we may conclude that for all $(i,j) \in [n_1] \times [n_2]$ the following bound is true:

$$p_{ij} \geq \frac{1}{(1+\alpha)^2}\hat{p}_{ij}. \tag{3.32}$$

For any $\epsilon \in (0,1)$ choosing m such that:

$$m = \frac{\log((n_1+n_2)\epsilon^{-1})}{2\left(\alpha \min_{i\in[n_1]} p_i^r \wedge \min_{j\in[n_2]} p_j^c\right)^2}, \tag{3.33}$$

guarantees that (3.32) holds with probability at least $(1-\epsilon)$ and the proof is complete. □

3.5.2 Proof of Lemma 3.4.2

Proof. The proof of Lemma 3.4.2 is similar to the previous proof but we include the full proof for completeness. We start our proof by analyzing empirical estimation of the row distribution; the analysis for the column distribution will be identical. Following the previous section we restrict ourselves to choose $\alpha \in (0, (\min_{i\in[n_1]} p_i^r \vee \min_{j\in[n_2]} p_j^c)^{-1})$. For any $i \in [n_1]$ choosing $t = \alpha \min_{i\in[n_1]} p_i^r$, by (3.25) we have that:

$$\Pr[|\hat{p}_i^r - p_i^r| \geq \alpha \min_{i\in[n_1]} p_i^r] \leq 2\exp(-2(\alpha \min_{i\in[n_1]} p_i^r)^2 m). \tag{3.34}$$

We may repeat the analysis for the column case, where we choose $t = \alpha \min_{j\in[n_2]} p_j^c$, then analogously:

$$\Pr[|\hat{p}_j^c - p_j^c| \geq \alpha \min_{j\in[n_2]} p_j^c] \leq 2\exp(-2(\alpha \min_{j\in[n_2]} p_j^c)^2 m). \tag{3.35}$$

For any $i \in [n_1]$ let E_i^r denote the event that $|\hat{p}_i^r - p_i^r| \geq \alpha \min_{i\in[n_1]} p_i^r$ and for any $j \in [n_2]$ let E_j^c denote the event that $|\hat{p}_j^c - p_j^c| \geq \alpha \min_{j\in[n_2]} p_j^c$. By (3.34), (3.35) and the Union Bound we have that:

$$\begin{aligned}
&\Pr\left[\text{For some } (i,j) \text{ the event } E_i^r \text{ or } E_j^c \text{ occurs}\right] && (3.36)\\
&\leq 2\left(n_1 \exp(-2(\alpha \min_{i\in[n_1]} p_i^r)^2 m) + n_2 \exp(-2(\alpha \min_{j\in[n_2]} p_j^c)^2 m\right)\\
&\leq 2(n_1+n_2)\exp(-2(\alpha \min_{i\in[n_1]} p_i^r \wedge \min_{j\in[n_2]} p_j^c)^2 m). && (3.37)
\end{aligned}$$

Observe that (3.37) immediately yields that with probability at least $1-2(n_1+n_2)\exp(-2(\alpha \min_{i\in[n_1]} p_i^r \wedge \min_{j\in[n_2]} p_j^c)^2 m)$ for any $(i,j) \in [n_1]\times[n_2]$ we have that the two following bounds hold:

$$|\hat{p}_i^r - p_i^r| \leq \alpha \min_{i\in[n_1]} p_i^r, \tag{3.38}$$

$$|\hat{p}_j^c - p_j^c| \leq \alpha \min_{j\in[n_2]} p_j^c. \tag{3.39}$$

The bound (3.38) is equivalent to the following:

$$-\alpha \min_{i\in[n_1]} p_i^r \leq \hat{p}_i^r - p_i^r \leq \alpha \min_{i\in[n_1]} p_i^r,$$

and the above inequality yields that for any $i \in [n_1]$:

$$\frac{1}{1+\alpha}\hat{p}_i^r \leq p_i^r \leq \frac{1}{1-\alpha}\hat{p}_i^r. \tag{3.40}$$

Similarly (3.39) implies that for any $j \in [n_2]$:

$$\frac{1}{1+\alpha}\hat{p}_j^c \leq p_j^c \leq \frac{1}{1-\alpha}\hat{p}_j^c. \tag{3.41}$$

Combining (3.40) and (3.41), we have that with probability at least $1 - 2(n_1 + n_2)\exp(-2(\alpha \min_{i\in[n_1]} p_i^r \wedge \min_{j\in[n_2]} p_j^c)^2 m)$ that:

$$\frac{1}{(1+\alpha)^2}\hat{\boldsymbol{p}} \leq \boldsymbol{p} \leq \frac{1}{(1-\alpha)^2}\hat{\boldsymbol{p}}. \tag{3.42}$$

For any $\epsilon \in (0,1)$ note that if we choose:

$$m = \frac{\log(2(n_1+n_2)\epsilon^{-1})}{2(\alpha \min_{i\in[n_1]} p_i^r \wedge \min_{j\in[n_2]} p_j^c)^2}, \tag{3.43}$$

then (3.42) holds with probability at least $1-\epsilon$ and the proof is complete. □

3.6 Matrix Completion Guarantees

With Lemma 3.4.1 in hand, we are prepared to prove Theorem 3.4.3 in Section 3.6.1. In Section 3.6.2, using Lemma 3.4.2 we will prove Theorem 3.4.4 which quantifies the relaxation of the condition for which (3.10) succeeds in obtaining exact recovery using the empirically learned weights when compared to unweighted nuclear norm minimization.

3.6.1 Proof of Theorem 3.4.3

Proof. For any $\alpha \in (0, (\min_{i\in[n]} p_i^r/(\sum_{i=1}^n p_i^r) \vee \min_{j\in[n]} p_j^c/(\sum_{j=1}^n p_j^c))^{-1})$ and $\epsilon \in (0,1)$ if we choose

$$m = \frac{1}{2}\left(\alpha \min_{i\in[n]} \frac{p_i^r}{\sum_{i=1}^n p_i^r} \wedge \min_{j\in[n]} \frac{p_j^c}{\sum_{j=1}^n p_j^c}\right)^{-2} \log(2\epsilon^{-1}n)$$

by Lemma 3.4.1 we have that with probability at least $(1-\epsilon)$ for any $(i,j) \in [n]\times[n]$:

$$\frac{p_{ij}}{\sum_{ij} p_{ij}} \geq \frac{1}{(1+\alpha)^2}\hat{p}_{ij}. \tag{3.44}$$

Observe that if the weights $\boldsymbol{R}, \boldsymbol{C}$ satisfy (3.16) for α, we have that:

$$p_{ij} \geq \frac{\sum_{ij} p_{ij}}{(1+\alpha)^2} \hat{p}_{ij} \tag{3.45}$$

$$\geq c_0 \left(\frac{R_i^2}{\sum_{i' \in \mathcal{S}_r} R_{i'}^2} + \frac{C_j^2}{\sum_{j' \in \mathcal{S}_c} C_{j'}^2} \right) \log^2(2n). \tag{3.46}$$

By Theorem 7 of [3], (3.46) is sufficient to guarantee exact recovery of $\boldsymbol{M}$ via (3.10) with probability at least $1 - 5(2n)^{-10}$. As stage one and stage two sampling are independent, we conclude that (3.10) attains exact recovery with probability at least $(1 - 5(2n)^{-10})(1 - \epsilon)$. □

3.6.2 Weighted Nuclear Norm and Relaxation of Sufficient Recovery Conditions

With Theorem 3.4.3 we established some sufficient conditions for the weights $\boldsymbol{R}, \boldsymbol{C}$ in order for (3.10) to attain exact recovery. In this section we will establish exact recovery guarantees for a specific set of weights defined in terms of the empirical sampling distribution $\hat{\boldsymbol{p}}$ and quantify how the exact recovery conditions for (3.10) are relaxed relative to unweighted nuclear norm minimization (3.4).

3.6.2.1 Proof of Theorem 3.4.4

Proof. Choosing the weights $\boldsymbol{R}, \boldsymbol{C}$ as in (3.17) and (3.18), observe that for any $(i, j) \in [n] \times [n]$:

$$\left(\frac{R_i^2}{\sum_{i' \in \mathcal{S}_r} R_{i'}^2} + \frac{C_j^2}{\sum_{j' \in \mathcal{S}_c} C_{j'}^2} \right) \log^2(2n) = \left(\frac{\hat{p}_i^r \sum_{j' \in \mathcal{S}_c} \hat{p}_{j'}^c + \hat{p}_j^c \sum_{i' \in \mathcal{S}_r} \hat{p}_{i'}^r}{\sum_{i', j' \in \mathcal{S}_r, \mathcal{S}_c} \hat{p}_{i'}^r \hat{p}_{j'}^c} \right) \log^2(2n). \tag{3.47}$$

Let $\alpha \in (0, (\min_{i\in[n]} p_i^r \vee \min_{j\in[n_2]} p_j^c)^{-1})$ be such that (3.19) and (3.20) hold and let $\epsilon \in (0,1)$ be arbitrary. By Lemma 3.4.2 choosing m such that:

$$m = \frac{1}{2}\left(\alpha \min_{i\in[n]} \frac{p_i^r}{\sum_{i=1}^n p_i^r} \wedge \min_{j\in[n]} \frac{p_j^c}{\sum_{j=1}^n p_j^c}\right)^{-2} \log(4\epsilon^{-1}n)$$

guarantees that with probability at least $(1-\epsilon)$ that for all indices $(i,j) \in [n]\times[n]$:

$$\frac{1}{(1+\alpha)^2}\hat{p}_{ij} \leq \frac{p_{ij}}{\sum_{i,j} p_{ij}} \leq \frac{1}{(1-\alpha)^2}\hat{p}_{ij}. \tag{3.48}$$

Applying (3.48) to (3.47) we have that for any $(i,j) \in [n]\times[n]$:

$$\begin{aligned}
\left(\frac{R_i^2}{\sum_{i'\in\mathcal{S}_r} R_{i'}^2} + \frac{C_j^2}{\sum_{j'\in\mathcal{S}_c} C_{j'}^2}\right)\log^2(2n) &= \left(\frac{\hat{p}_i^r \sum_{j'\in\mathcal{S}_c}\hat{p}_{j'}^c + \hat{p}_j^c\sum_{i'\in\mathcal{S}_r}\hat{p}_{i'}^r}{\sum_{i',j'\in\mathcal{S}_r,\mathcal{S}_c}\hat{p}_{i'}^r\hat{p}_{j'}^c}\right)\log^2(2n) \\
&\leq \frac{(1+\alpha)^2}{(1-\alpha)^2}\left(\frac{p_i^r\sum_{j'\in\mathcal{S}_c}p_{j'}^c + p_j^c\sum_{i'\in\mathcal{S}_r}p_{i'}^r}{\sum_{i',j'\in\mathcal{S}_r,\mathcal{S}_c}p_{i'}^r p_{j'}^c}\right)\log^2(2n) \\
&= \frac{(1+\alpha)^2}{(1-\alpha)^2}\left[\frac{p_i^r\log^2(2n)}{\sum_{i'\in\mathcal{S}_r}p_{i'}^r} + \frac{p_j^c\log^2(2n)}{\sum_{j'\in\mathcal{S}_c}p_{j'}^c}\right] \\
&\leq \frac{(1+\alpha)^2}{(1-\alpha)^2}\left[\frac{p_i^r\log^2(2n)}{\sum_{i'\in\mathcal{S}_r^*}p_{i'}^r} + \frac{p_j^c\log^2(2n)}{\sum_{j'\in\mathcal{S}_c^*}p_{j'}^c}\right] \qquad (3.49) \\
&\leq \frac{1}{c_0}p_{ij}. \qquad (3.50)
\end{aligned}$$

where (3.49) follows as the sets $\mathcal{S}_r^*, \mathcal{S}_c^*$ serve as a lower bound for the terms $\sum_{i'\in\mathcal{S}_r} p_{i'}^r, \sum_{j'\in\mathcal{S}_c} p_{j'}^c$ respectively and thus inverting they serve as an upper bound and (3.50) follows from (3.19) and (3.20). Again by Theorem 7 of [3] we immediately see that (3.50) is sufficient to guarantee exact recovery of $\boldsymbol{M}$ via (3.10) with probability at least $1 - 5(2n)^{-10}$. □

3.7 Numerical Experiments

Here we test the performance of weighted nuclear norm minimization using various weights. We have the following experimental setup: the data matrix $\boldsymbol{M}$ is a unit Frobenius norm standard normal Gaussian square matrix of dimension $n = 500$. Our sampling distribution $\boldsymbol{p} = \boldsymbol{p^r p^c}$ where $\boldsymbol{p^r}, \boldsymbol{p^c}$ are power law distributed with exponent equal to 1.2. Sampling the distribution $\boldsymbol{p}$ at a rate of m times with replacement and we obtain the empirical distribution $\hat{\boldsymbol{p}} = \hat{\boldsymbol{p}}^{\boldsymbol{r}}\hat{\boldsymbol{p}}^{\boldsymbol{c}}$. Using this empirical distribution $\hat{\boldsymbol{p}}$ we test nuclear norm minimization using the following weights, as was done in [32]:

1. Unweighted (Uniform Weights): the weights $\boldsymbol{R}, \boldsymbol{C}$ are equal to the uniform weights.

2. True Weighted: the weights $\boldsymbol{R}, \boldsymbol{C}$ satisfy: $\boldsymbol{R} = (\boldsymbol{p^r})^{1/2}, \boldsymbol{C} = (\boldsymbol{p^c})^{1/2}$.

3. Empirically Weighted: the weights $\boldsymbol{R}, \boldsymbol{C}$ satisfy: $\boldsymbol{R} = (\hat{\boldsymbol{p}}^{\boldsymbol{r}})^{1/2}, \boldsymbol{C} = (\hat{\boldsymbol{p}}^{\boldsymbol{c}})^{1/2}$.

4. Empirically Smoothed Weights: the weights $\boldsymbol{R}, \boldsymbol{C}$ are a linear combination of the empirical weights and the uniform weights. Letting $\mathbf{1}_n := [1, \ldots, 1]$ be a vector of length n whose coordinates are all equal to 1, we set $\boldsymbol{R} = \frac{1}{2n}(\hat{\boldsymbol{p}}^{\boldsymbol{r}})^{1/2} + \frac{1}{2n}\mathbf{1}_n$ and $\boldsymbol{C} = \frac{1}{2}(\hat{\boldsymbol{p}}^{\boldsymbol{c}})^{1/2} + \frac{1}{2n}\mathbf{1}_n$, i.e. we put half of the mass on the empirical distribution and remaining half of the mass on the uniform weights.

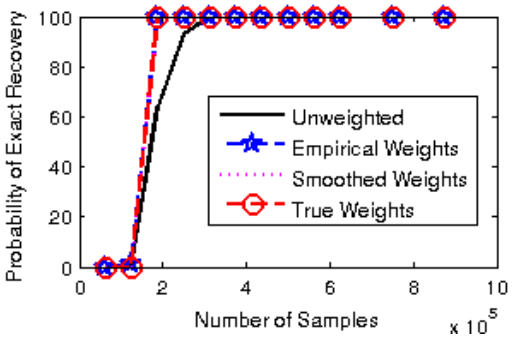

Figure 3.1: Probability of Exact Recovery when the rank is equal to 5.

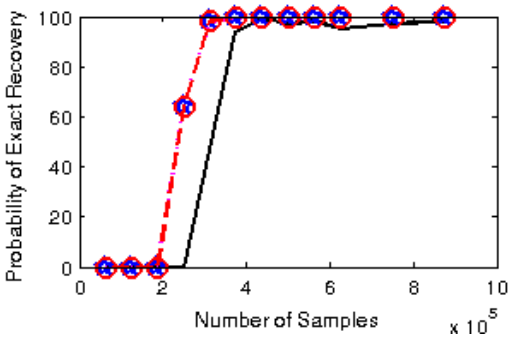

Figure 3.2: Probability of Exact Recovery when the rank is equal to 10.

We let the rank of $\boldsymbol{M}$ be 5, 10, 15, 20, 25 and we choose a range of variable sampling rates. For each rank and sampling rate test configuration we performed 100 trials. We consider exact recovery to be when the output of the weighted nuclear norm $\bar{\boldsymbol{M}}$ satisfies: $\|\boldsymbol{M} - \bar{\boldsymbol{M}}\|_F \leq 10^{-5}$. To execute the weighted nuclear norm minimization program we utilized the Augmented Lagrangian Method [34]. We obtained the following phase transition diagrams in Figures 3.1-3.5.

Note that we do not perform the two stage sampling method. As the power law sampling distribution $\boldsymbol{p}$ is non-uniform, even though we may sample at a rate of $m = O(n_1 n_2)$, the rate that the percentage of unique revealed entries of $\boldsymbol{M}$ grows is in line with the uniform sampling regime we are accustomed to. In Figure 3.6 we show how with the independent sampling with replacement rate m grows with the percentage of unique entries of $\boldsymbol{M}$.

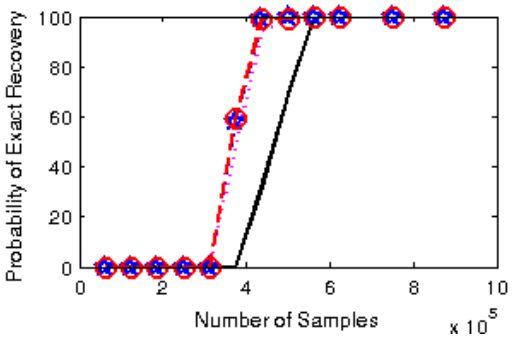

Figure 3.3: Probability of Exact Recovery when the rank is equal to 15.

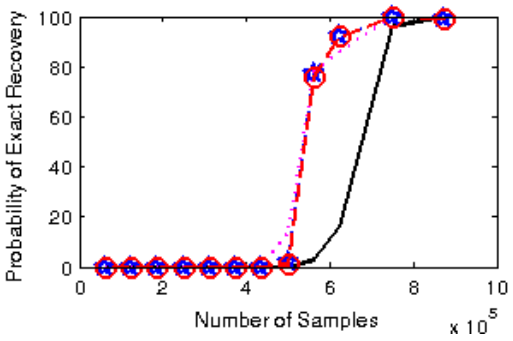

Figure 3.4: Probability of Exact Recovery when the rank is equal to 20.

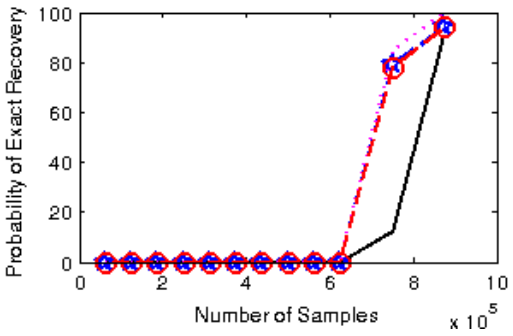

Figure 3.5: Probability of Exact Recovery when the rank is equal to 25.

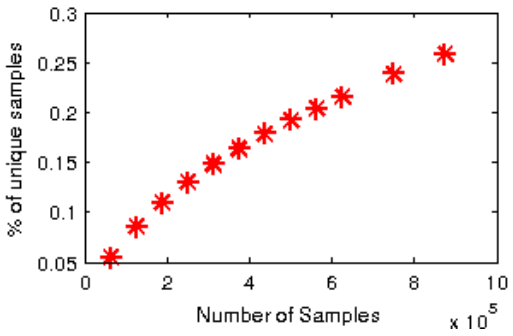

Figure 3.6: Power Law Sampling with replacement rate vs. Percentage of Unique Samples Revealed.

3.8 Conclusion

We extended two weighted nuclear norm minimization results from [3]. In particular we extended results where the weights were being defined in relation to the true sampling distribution $\boldsymbol{p}$ to the weights being defined in relation to the empirical sampling distribution $\hat{\boldsymbol{p}}$. Furthermore, we defined an empirical set of weights and established a quantifiable relaxation of exact

recovery conditions for weighted nuclear norm minimization when compared to the unweighted nuclear norm. To achieve these guarantees we used a large deviation bound and a concentration of measure inequality from [33]. We showed that weighted nuclear norm minimization is quite robust to the choice of empirically learned weights. Indeed, we used a broad range of empirical weights and saw strikingly similar exact recovery gains over unweighted nuclear norm minimization.

Chapter 4

Alternate Weighted Matrix Completion Analysis

4.1 Introduction

In Chapter 3 we provided weighted nuclear norm minimization guarantees using concentration of measure and large deviation bounds from [33]. In this chapter we shall present an alternate analysis of these guarantees using tools from the theory of convergence of empirical processes [35–38]. Roughly speaking, note that for both Theorems 3.4.3 and 3.4.4 that the sampling bound m depends inverse quadratically on the true sampling distribution $\boldsymbol{p}$ and has an ϵ^{-1} logarithmic dependence. In this chapter, we use a result on the convergence of fluctuation processes to obtain new sampling bounds for m. To this end, we derive sampling bounds m in which the bounds m no longer depend on the true sampling distribution $\boldsymbol{p}$ in an inverse quadratic fashion. The trade-off is that we exchange a logarithmic dependence on ϵ^{-1} with a logarithmic dependence on ϵ^{-2} and that our sampling bound depends on the convergence rate of the fluctuation process. We believe that these sampling bounds may be of independent interest than those bounds obtained in Chapter 3 depending on the sampling distribution $\boldsymbol{p}$. For example, if the sampling distribution is extremely non-uniform, perhaps the inverse quadratic dependence on $\boldsymbol{p}$ would

require too many samples and instead, in the analysis, one would prefer a logarithmic dependence on $\boldsymbol{p}$.

4.2 Main Results

As in Chapter 3 we shall assume that our sampling distribution $\boldsymbol{p}$ has a product form $p_{ij} = p_i^r p_j^c$ for all $(i, j) \in [n_1] \times [n_2]$ and we will consider the following *two-stage sampling model*:

- Stage 1 (Empirical Sampling Distribution): We sample the distribution $\boldsymbol{p}$ with m times independently with replacement, but the corresponding entries of the data matrix $\boldsymbol{M}$ are not revealed to us. In other words, we are *sampling the sampling distribution*, but not the underlying matrix $\boldsymbol{M}$.

- Stage 2 (Sampling the Matrix): We then, independent of the first stage, sample the matrix $\boldsymbol{M}$ using the independent Bernoulli model for each entry $(i, j) \in [n_1] \times [n_2]$.

We are now ready to state our main results in the context of this two stage sampling model. In the initial sampling stage we are only sampling the distribution $\boldsymbol{p}$ and we are not sampling the actual entries of the matrix $\boldsymbol{M}$, merely its indices. Therefore, during this stage of sampling, our goal is to obtain sufficient sampling lower bounds in order to guarantee that with arbitrary probability, our empirical distribution $\hat{\boldsymbol{p}}$ will serve as a valid proxy for $\boldsymbol{p}$ in

Theorem 3.3.2. We use the following notation: for two vectors $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{x} \leq \boldsymbol{y}$ denotes an element-wise inequality, i.e. $x_i \leq y_i$ for all i; similarly for other order statements involving vectors or arrays. To this end, we prove the following result for the empirical estimation of product distributions:

Lemma 4.2.1. *Let $\boldsymbol{p} = \boldsymbol{p^r}\boldsymbol{p^c}$ where $\boldsymbol{p^r}, \boldsymbol{p^c}$ are probability mass functions on $[n_1], [n_2]$ respectively. For any $1/\beta, \epsilon \in (0,1)$ there exists an index $m_\epsilon \in \mathbb{N}$ and a constant c_ϵ such that for all $(i,j) \in [n_1] \times [n_2]$, we have that:*

$$\Pr\left[p_{ij} \geq \frac{1}{\beta}\hat{p}_{ij}\right] \geq (1-\epsilon)^2, \tag{4.1}$$

provided that the number of stage one samples $m = \left(\frac{c_\epsilon}{1-\sqrt{\beta}}\right)^2$ and:

$$c_\epsilon \geq O\left(1 \wedge \log^{\frac{1}{2}}\left(\epsilon^{-2}(2\pi)^{1-n}\left(\min\left(\Pi_{i=1}^{n_1} p_i^r, \Pi_{j=1}^{n_2} p_j^c\right)\right)^{-1}\right) \wedge \sqrt{m_\epsilon}|1-\sqrt{\beta}|\right), \tag{4.2}$$

where $a \wedge b := \max(a,b)$.

Note that Lemma 4.2.1 is a general result for the empirical estimation of any distribution $\boldsymbol{p}$ over $[n_1] \times [n_2]$ which has the following product form: $p_{ij} = p_i^r p_j^c$ where $\boldsymbol{p^r}, \boldsymbol{p^c}$ are probability distributions on $[n_1], [n_2]$ respectively. Recall that the sampling model employed in [3] is a sequence of $n_1 \cdot n_2$ independent Bernoulli random variables, with each Bernoulli random variable having success probability p_{ij} for $(i,j) \in [n_1] \times [n_2]$. Therefore, $\boldsymbol{p}$ may not be a probability distribution as it may not sum to 1. To this end, we note that when we sample $\boldsymbol{p}$, we are really sampling the normalized matrix $\frac{1}{\sum_{i,j} p_{ij}}\boldsymbol{p}$. So our empirical estimator $\hat{\boldsymbol{p}}$ is estimating the normalized matrix $\frac{1}{\sum_{i,j} p_{ij}}\boldsymbol{p}$ and

not $\boldsymbol{p}$ itself. Therefore, in order to apply Lemma 4.2.1 we must account for this normalization constant.

Having noted the above, with Lemma 4.2.1 in hand we may pass from the weights $\boldsymbol{R}, \boldsymbol{C}$ being defined in terms of the true sampling distribution $\boldsymbol{p}$ to being defined in terms of the empirical sampling distribution $\hat{\boldsymbol{p}}$ to obtain the following weighted nuclear norm minimization guarantee:

Theorem 4.2.2. *Let $\boldsymbol{M} = (M_{ij})$ be an $n \times n$ matrix of rank-r, and suppose that its elements M_{ij} are observed only over a subset of elements $\Omega \subset [n] \times [n]$. Let $\epsilon \in (0, 1)$ be arbitrary. Suppose that there exists some $\beta > 1$ and some universal constant c_0 such that for all indices $(i, j) \in [n] \times [n]$ the weights $\boldsymbol{R}, \boldsymbol{C}$ satisfy the following inequalities:*

$$\hat{p}_{ij} \geq \frac{\beta c_0}{\sum_{i,j} p_{ij}} \left(\frac{R_i^2}{\sum_{i' \in \mathcal{S}_r} R_{i'}^2} + \frac{C_j^2}{\sum_{j' \in \mathcal{S}_c} C_{j'}^2} \right) \log^2(2n). \tag{4.3}$$

If the parameters c_ϵ, m are chosen as in Lemma 4.2.1 and if for all $(i, j) \in [n] \times [n], p_{ij} \geq n^{-10}$, then with probability at least $(1 - 5(2n)^{-10})(1 - \epsilon)^2$, $\boldsymbol{M}$ is the unique optimum to (3.10), where Ω in (3.10) is obtained via the usual (stage two) independent entry-wise Bernoulli sampling of $\boldsymbol{M}$.

We also establish the following *two sided empirical bound* for the estimation of product distributions:

Lemma 4.2.3. *Let $\boldsymbol{p} = \boldsymbol{p^r p^c}$ where $\boldsymbol{p^r}, \boldsymbol{p^c}$ are probability mass functions on $[n_1], [n_2]$ respectively. Let $1/\beta_1, 1/\beta_2, \epsilon \in (0, 1)$ be arbitrary, with $1/\beta_1 \leq 1/\beta_2$.*

Then there exists an index m_ϵ and constants $c_{\epsilon,1}, c_{\epsilon,2}$ such that for all $(i,j) \in [n_1] \times [n_2]$:

$$\Pr\left[\frac{1}{\beta_1}\hat{p}_{ij} \leq p_{ij} \leq \frac{1}{\beta_2}\hat{p}_{ij}\right] \geq (1-\epsilon)^2, \tag{4.4}$$

provided that we choose:

$$m = \left(\frac{c_{\epsilon,2}}{1-\sqrt{\beta_2}}\right)^2 \tag{4.5}$$

$$c_{\epsilon,1} = -\left|\frac{1-\sqrt{\beta_1}}{1-\sqrt{\beta_2}}\right| c_{\epsilon,2}, \tag{4.6}$$

$$c_{\epsilon,2} \geq O\left(1 \wedge \log^{\frac{1}{2}}\left(\epsilon^{-2}(2\pi)^{1-n}\left(\min\left(\Pi_{i=1}^{n_1} p_i^r, \Pi_{j=1}^{n_2} p_j^c\right)\right)^{-1} \wedge \sqrt{m_\epsilon}|1-\sqrt{\beta_2}|\right)\right). \tag{4.7}$$

Using Lemma 4.2.3, we show that for a specific set of empirically defined weights $\boldsymbol{R}, \boldsymbol{C}$ we may quantify the relaxation of the exact recovery conditions for (3.10) versus (3.4):

Theorem 4.2.4. *Let $\boldsymbol{M}$ be a square $n \times n$ rank-r matrix with coherence μ_0. Consider the weights defined by:*

$$R_i = \sqrt{\frac{1}{n}\hat{p}_i^r \sum_{j' \in \mathcal{S}_c} \hat{p}_{j'}^c} \text{ for } i = 1, \ldots, n, \tag{4.8}$$

$$C_j = \sqrt{\frac{1}{n}\hat{p}_j^c \sum_{i' \in \mathcal{S}_r} \hat{p}_{i'}^r} \text{ for } j = 1, \ldots, n, \tag{4.9}$$

where $\mathcal{S}_r, \mathcal{S}_c$ denote the $\lfloor n/(\mu_0 r)\rfloor$ entries of $\hat{\boldsymbol{p}}^{\boldsymbol{r}}, \hat{\boldsymbol{p}}^{\boldsymbol{c}}$ of least magnitude, respectively. Let the sets $\mathcal{S}_r^*, \mathcal{S}_c^*$ denote the $\lfloor n/(\mu_0 r)\rfloor$ entries of $\boldsymbol{p}^{\boldsymbol{r}}, \boldsymbol{p}^{\boldsymbol{c}}$ of least magnitude, respectively. Suppose that there exists a constant $\mathcal{C} \leq 1/4$ such

that the (unnormalized) matrix $\boldsymbol{p}$ *satisfies for all* $(i,j) \in [n] \times [n]$*:*

$$p_j^c \sum_{i' \in \mathcal{S}_r^*} p_{i'}^r \geq \frac{c_0}{\mathcal{C}} \log^2(2n), \tag{4.10}$$

$$p_i^r \sum_{j' \in \mathcal{S}_c^*} p_{j'}^c \geq \frac{c_0}{\mathcal{C}} \log^2(2n). \tag{4.11}$$

Then for any $\epsilon \in (0,1)$*, there exists an index* $m_\epsilon \in \mathbb{N}$ *and* β_1, β_2 *such that if* $c_{\epsilon,1}, c_{\epsilon,2}, m$ *are chosen as in* (4.5)-(4.7) *and if for all* $(i,j) \in [n] \times [n], p_{ij} \geq \frac{1}{n^{10}}$ *then* (3.10) *achieves exact recovery with probability at least* $(1 - 5(2n)^{-10})(1 - \epsilon)^2$*, where* Ω *in* (3.10) *is chosen via the (stage two) independent entry-wise Bernoulli sampling of* $\boldsymbol{M}$*.*

Note: Unweighted nuclear norm minimization attains exact recovery under the condition that for all $(i,j) \in [n] \times [n]$:

$$p_i^r p_j^c \gtrsim \frac{\mu_0 r}{n} \log^2(2n). \tag{4.12}$$

However as Theorem 4.2.4 establishes, weighted nuclear norm minimization with choice of weights (4.8) and (4.9) attains exact recovery subject to the less restrictive sufficient recovery condition:

$$p_j^c \sum_{i' \in \mathcal{S}_r^*} p_{i'}^r \gtrsim \log^2(2n),$$

$$p_i^r \sum_{j' \in \mathcal{S}_c^*} p_{j'}^c \gtrsim \log^2(2n),$$

precisely the condition from [3].

4.3 Empirical Estimation

Given any probability mass function $\boldsymbol{p}$ on $[n]$, we will sample this distribution with replacement m times. Given the $X_1, \ldots, X_m \overset{i.i.d}{\sim} \boldsymbol{p}$ random variables, we may define the *empirical estimator*:

Definition 4.3.1. The empirical estimator $\hat{\boldsymbol{p}}$ is defined by:

$$\hat{p}_i := \frac{1}{m} \sum_{j=1}^{m} \delta_i(X_j), \text{ for } i \in [n]. \tag{4.13}$$

Definition 4.3.2. For any probability mass function $\boldsymbol{p}$ on $[n]$ we may define the fluctuation process $\boldsymbol{Y}_m$ by:

$$Y_{m,i} = \sqrt{m}(\hat{p}_i - p_i), \text{ for } i \in [n]. \tag{4.14}$$

From [39], we have the following empirical convergence result:

Theorem 4.3.1. *The fluctuation process $\boldsymbol{Y}_m$ converges in distribution to $\boldsymbol{Y}$, a multivariate mean zero Gaussian random vector with covariance matrix $\boldsymbol{\Sigma}$ given by:*

$$\Sigma_{ij} = \begin{cases} -p_i p_j & \text{if } i \neq j \\ (1-p_i)p_i & \text{if } i = j \end{cases}. \tag{4.15}$$

For further background on the theory of empirical processes and their convergence, we refer the reader to [35–38].

For ease of notation, we allow c to denote c_ϵ and c_1, c_2 to denote $c_{\epsilon,1}, c_{\epsilon,2}$ respectively for the remainder of the chapter.

4.3.1 Proof of Lemma 4.2.1

Proof. With all the statistical preliminaries out of the way, we are now prepared to prove Lemma 4.2.1. Since we are assuming that $p_{ij} = p_i^r p_j^c$, it suffices to obtain the following inequalities:

$$\begin{aligned} \Pr\left[\boldsymbol{p^r} \geq \frac{1}{\sqrt{\beta}}\hat{\boldsymbol{p}}^{\boldsymbol{r}}\right] &\geq (1-\epsilon), \\ \Pr\left[\boldsymbol{p^c} \geq \frac{1}{\sqrt{\beta}}\hat{\boldsymbol{p}}^{\boldsymbol{c}}\right] &\geq (1-\epsilon). \end{aligned} \tag{4.16}$$

Since the distributions $\boldsymbol{p^r}$ and $\boldsymbol{p^c}$ are independent, if the above holds, then we may conclude that for all $(i,j) \in [n_1] \times [n_2]$ that $\Pr\left[p_{ij} \geq \frac{1}{\beta}\hat{p}_{ij}\right] \geq (1-\epsilon)^2$ as desired. Therefore, without loss of generality, for ease of notation's sake, it suffices to consider $\boldsymbol{p} = \boldsymbol{p^r}$ and $\hat{\boldsymbol{p}} = \hat{\boldsymbol{p}}^{\boldsymbol{r}}$; the exact same analysis holds for the column distribution $\boldsymbol{p^c}$ and its empirical estimator $\hat{\boldsymbol{p}}^{\boldsymbol{c}}$.

Observe that if $\sqrt{m}\,(\hat{\boldsymbol{p}} - \boldsymbol{p}) \xrightarrow{\mathcal{D}} \boldsymbol{Y}$, then $\sqrt{m}\,(\boldsymbol{p} - \hat{\boldsymbol{p}}) \xrightarrow{\mathcal{D}} -\boldsymbol{Y}$ and since $\boldsymbol{Y}$ is a mean zero multivariate Gaussian, $\boldsymbol{Y}$ and $-\boldsymbol{Y}$ have identical probability distribution functions due to symmetry.

Consider the term $\Pr[\sqrt{m}(\boldsymbol{p} - \hat{\boldsymbol{p}}) \geq c\boldsymbol{p}]$. We have that:

$$\Pr[\sqrt{m}(\boldsymbol{p} - \hat{\boldsymbol{p}}) \geq c\boldsymbol{p}] = \Pr\left[\boldsymbol{p} \geq \frac{1}{\left(1 - \frac{c}{\sqrt{m}}\right)}\hat{\boldsymbol{p}}\right], \tag{4.17}$$

precisely the type of inequality we desire. Note that we require that $\frac{1}{\left(1-\frac{c}{\sqrt{m}}\right)} \in (0,1)$ *not* that $c \in (0,1)$. In fact from the minus sign in (4.17) we would have to require $c < 0$, else we violate the fact that $\boldsymbol{p}, \hat{\boldsymbol{p}}$ are probability distributions. For the remainder of this proof we will adopt the convention that the constant

$c > 0$. Employing this convention, (4.17) equivalently becomes:

$$\Pr[\sqrt{m}(\boldsymbol{p} - \hat{\boldsymbol{p}}) \geq -c\boldsymbol{p}] = \Pr\left[\boldsymbol{p} \geq \frac{1}{\left(1 + \frac{c}{\sqrt{m}}\right)}\hat{\boldsymbol{p}}\right]. \tag{4.18}$$

Due to convergence in distribution, we also have that:

$$\lim_{m\to\infty} \Pr[\sqrt{m}(\boldsymbol{p} - \hat{\boldsymbol{p}}) \geq -c\boldsymbol{p}] = \Pr[-\boldsymbol{Y} \geq -c\boldsymbol{p}] = \Pr[\boldsymbol{Y} \geq -c\boldsymbol{p}]. \tag{4.19}$$

By (4.19) we may conclude that for any $\epsilon \in (0, 1)$, there exists an index m_ϵ such that for all $m \geq m_\epsilon$:

$$\left|\Pr\left[\boldsymbol{p} \geq \frac{1}{\left(1 + \frac{c}{\sqrt{m}}\right)}\hat{\boldsymbol{p}}\right] - \Pr[\boldsymbol{Y} \geq -c\boldsymbol{p}]\right| \leq \epsilon/2, \tag{4.20}$$

which implies that:

$$\Pr\left[\boldsymbol{p} \geq \frac{1}{\left(1 + \frac{c}{\sqrt{m}}\right)}\hat{\boldsymbol{p}}\right] \geq \Pr[\boldsymbol{Y} \geq -c\boldsymbol{p}] - \epsilon/2 \tag{4.21}$$

Therefore, it suffices to obtain the bound:

$$\Pr[\boldsymbol{Y} \geq -c\boldsymbol{p}] \geq 1 - \epsilon/2. \tag{4.22}$$

We have that $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is defined in (4.15). However by [40], the multinomial covariance matrix $\boldsymbol{\Sigma}$ is not full rank and therefore $\boldsymbol{Y}$ has a degenerate probability distribution function. In particular, $\boldsymbol{\Sigma}$ has rank $n - 1$. When a covariance matrix has rank $n - 1$ then with probability 1, one of the coordinates Y_i is a linear combination of the remaining coordinates

$\{Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n\}$. Without loss of generality, we may assume that there exists scalars $\{\alpha_i\}_{i=1}^{n-1}$ such that:

$$Y_n = \sum_{i=1}^{n-1} \alpha_i Y_i. \tag{4.23}$$

Observe that if $Y_i \geq -cp_i$ for all $i = 1, \ldots, n-1$, then we have that:

$$\begin{aligned} Y_n &= \sum_{i=1}^{n-1} \alpha_i Y_i \\ &\geq \sum_{i=1}^{n-1} -\alpha_i c p_i \\ &\geq -c(\max_i(|\alpha_i|, 1)) \sum_{i=1}^{n-1} p_i \\ &= -c(\max_i(|\alpha_i|, 1))(1 - p_n) \\ &= -\tilde{c} p_n, \end{aligned}$$

for some new constant $\tilde{c}$ and $\tilde{c} \geq c > 0$.

For improved legibility, we let $\boldsymbol{Y}_{n-1} := (Y_1, \ldots, Y_{n-1})$ and $\boldsymbol{p}_{n-1} := (p_1, \ldots, p_{n-1})$. Let $\tilde{c}_{\epsilon/2}$ denote the constant such that:

$$\Pr[\boldsymbol{Y}_{n-1} \geq -\tilde{c}_{\epsilon/2} \boldsymbol{p}_{n-1}] = 1 - \epsilon/2.$$

With abuse of notation, setting $c = \max(\tilde{c}_{\epsilon/2}, c)$ yields that $\Pr[(Y_1, \ldots, Y_n) \geq -c(p_1, \ldots, p_n)] \geq 1 - \epsilon/2$. Having eliminated the last coordinate Y_n from our analysis, we have that $\boldsymbol{Y}_{n-1} \sim \mathcal{N}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}})$ where $\hat{\boldsymbol{\Sigma}}$ is the principle $(n-1) \times (n-1)$ submatrix of $\boldsymbol{\Sigma}$, i.e. $\hat{\boldsymbol{\Sigma}}_{i,j} = (\boldsymbol{\Sigma})_{i,j}$ for $1 \leq i, j \leq n-1$. The key fact here is

that $\hat{\boldsymbol{\Sigma}}$ is full rank $n-1$ and by [40] its inverse is given by:

$$\hat{\boldsymbol{\Sigma}}^{-1} = \frac{1}{p_n} \begin{bmatrix} \frac{p_1+p_n}{p_1} & 1 & \cdots & 1 \\ 1 & \frac{p_2+p_n}{p_2} & 1 & \vdots \\ \vdots & 1 & \ddots & 1 \\ 1 & \cdots & 1 & \frac{p_{n-1}+p_n}{p_{n-1}} \end{bmatrix}. \tag{4.24}$$

Therefore the quadratic term $\boldsymbol{x}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{x}$ has the following form:

$$\begin{aligned} \boldsymbol{x}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{x} &= \frac{1}{p_n} \left[2 \sum_{i \neq j} x_i x_j + \sum_{i=1}^{n-1} x_i^2 \frac{p_i + p_n}{p_i} \right] \\ &= \frac{2}{p_n} \sum_{i \neq j} x_i x_j + \sum_{i=1}^{n-1} x_i^2 \left(\frac{p_i + p_n}{p_i p_n} \right) \\ &= \frac{2}{p_n} \sum_{i \neq j} x_i x_j + \frac{1}{p_n} \sum_{i=1}^{n-1} x_i^2 + \sum_{i=1}^{n-1} \frac{x_i^2}{p_i} \\ &= \frac{1}{p_n} \left(x_1 + \cdots + x_{n-1} \right)^2 + \sum_{i=1}^{n-1} \frac{x_i^2}{p_i}. \end{aligned} \tag{4.25}$$

Therefore we obtain:

$$\begin{aligned} &\Pr[\boldsymbol{Y}_{n-1} \geq -c\boldsymbol{p}_{n-1}] \\ &= \frac{1}{\sqrt{(2\pi)^{n-1} |\hat{\boldsymbol{\Sigma}}|}} \int_{-cp_1}^{\infty} \cdots \int_{-cp_{n-1}}^{\infty} \exp\left(-\frac{1}{2} \boldsymbol{x}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{x} \right) d\boldsymbol{x}_{n-1} \\ &= \frac{1}{\sqrt{(2\pi)^{n-1} |\hat{\boldsymbol{\Sigma}}|}} \int_{-cp_1}^{\infty} \cdots \int_{-cp_{n-1}}^{\infty} \exp\left(-\frac{1}{2} \left(\frac{1}{p_n} \left(x_1 + \cdots + x_{n-1} \right)^2 + \sum_{i=1}^{n-1} \frac{x_i^2}{p_i} \right) \right) d\boldsymbol{x}_{n-1}, \end{aligned}$$

where $d\boldsymbol{x}_{n-1} := dx_1 \cdots dx_{n-1}$.

We now turn our attention to bounding the integral I defined by:

$$I := \frac{1}{\sqrt{(2\pi)^{n-1} |\hat{\boldsymbol{\Sigma}}|}} \int_{-cp_1}^{\infty} \cdots \int_{-cp_{n-1}}^{\infty} \exp\left(-\frac{1}{2} \left(\frac{1}{p_n} \left(x_1 + \cdots + x_{n-1} \right)^2 + \sum_{i=1}^{n-1} \frac{x_i^2}{p_i} \right) \right) d\boldsymbol{x}_{n-1}.$$

We now note that by symmetry of the normal distribution that finding a constant $c > 0$ such that $I \geq 1 - \epsilon/2$ is equivalent to finding a constant $c > 0$ such that $I^+ \leq \epsilon/2$, where:

$$I^+ := \frac{1}{\sqrt{(2\pi)^{n-1}|\hat{\boldsymbol{\Sigma}}|}} \int_{cp_1}^{\infty} \cdots \int_{cp_{n-1}}^{\infty} \exp\left(-\frac{1}{2}\left(\frac{1}{p_n}\left(x_1 + \cdots + x_{n-1}\right)^2 + \sum_{i=1}^{n-1} \frac{x_i^2}{p_i}\right)\right) d\boldsymbol{x}_{n-1}. \tag{4.26}$$

For the remainder of this proof, we will be considering this case.

When $c = 1$, we make the following definition:

$$\epsilon_1 := \frac{1}{\sqrt{(2\pi)^{n-1}|\hat{\boldsymbol{\Sigma}}|}} \int_{p_1}^{\infty} \cdots \int_{p_{n-1}}^{\infty} \exp\left(-\frac{1}{2}\left(\frac{1}{p_n}\left(x_1 + \cdots + x_{n-1}\right)^2 + \sum_{i=1}^{n-1} \frac{x_i^2}{p_i}\right)\right) d\boldsymbol{x}_{n-1}. \tag{4.27}$$

Hence if $\epsilon/2 \geq \epsilon_1$, it suffices to choose $c = 1$ as our constant. The remainder of the proof is concerned with the case when $\epsilon/2 < \epsilon_1$ or equivalently when $c > 1$.

Observe that on the integration domain $D = \{(x_1, \ldots, x_{n-1}) \in \mathbb{R}^{n-1} : x_i \geq cp_i, \text{ for } i = 1, \ldots, n-1\}$ we have that:

$$\exp\left(-\frac{1}{2p_n}(x_1 + \cdots + x_{n-1})^2\right) \leq \exp\left(-\frac{1}{2p_n}c^2(1 - p_n)^2\right),$$

thus yielding that:

$$I^+ \leq \frac{\exp\left(-\frac{1}{2p_n}c^2(1-p_n)^2\right)}{\sqrt{(2\pi)^{n-1}|\hat{\boldsymbol{\Sigma}}|}} \prod_{i=1}^{n-1} \int_{cp_i}^{\infty} \exp\left(-\frac{1}{p_i}x_i^2\right) dx_i.$$

Now each integral $I_i := \int_{cp_i}^{\infty} \exp\left(-\frac{1}{p_i}x_i^2\right) dx_i$ can be bounded in the following

manner:

$$\begin{aligned}
I_i &\leq \int_{cp_i}^{\infty} \frac{x_i}{cp_i} \exp\left(-\frac{1}{2p_i}x_i^2\right) dx_i \\
&= \frac{1}{cp_i}\int_{cp_i}^{\infty} x_i \exp\left(-\frac{1}{2p_i}x_i^2\right) dx_i \\
&= \frac{1}{cp_i}\int_{cp_i}^{\infty} x_i \exp\left(-\left(\frac{x_i}{\sqrt{2p_i}}\right)^2\right) dx_i \\
&= \frac{1}{cp_i}\int_{c\sqrt{p_i/2}}^{\infty} u\sqrt{2p_i}\exp(-u^2)\sqrt{2p_i}du \\
&= \frac{2}{c}\int_{c\sqrt{p_i/2}}^{\infty} u\exp(-u^2)du \\
&= \frac{1}{c}\exp(-c^2p_i/2).
\end{aligned}$$

We therefore have that:

$$\begin{aligned}
I^+ &\leq \frac{\exp\left(-\frac{1}{2p_n}c^2(1-p_n)^2\right)}{\sqrt{(2\pi)^{n-1}|\hat{\boldsymbol{\Sigma}}|}}\frac{1}{c^{n-1}}\exp\left(-\frac{c^2}{2}(1-p_n)\right) \\
&= \frac{\exp\left(-\frac{1}{2}c^2(1-p_n)\left[(1-p_n)/p_n+1\right]\right)}{c^{n-1}\sqrt{(2\pi)^{n-1}|\hat{\boldsymbol{\Sigma}}|}} \\
&= \frac{\exp\left(-\frac{1}{2p_n}c^2(1-p_n)\right)}{c^{n-1}\sqrt{(2\pi)^{n-1}|\hat{\boldsymbol{\Sigma}}|}} \\
&\leq \frac{\exp\left(-\frac{1}{2p_n}c^2(1-p_n)\right)}{\sqrt{(2\pi)^{n-1}|\hat{\boldsymbol{\Sigma}}|}} \qquad (4.28)
\end{aligned}$$

Where (4.28) holds since $c > 1$. We now wish to bound (4.28) above by $\epsilon/2$. Setting the RHS of (4.28) less than or equal to $\epsilon/2$ and solving for c yields:

$$c \geq O\left(\log^{\frac{1}{2}}\left(\epsilon^{-2}(2\pi)^{1-n}|\hat{\boldsymbol{\Sigma}}|^{-1}\right)\right) \qquad (4.29)$$

Thus, we see that:

$$c \geq O\left(1 \wedge \log^{\frac{1}{2}}\left(\epsilon^{-2}(2\pi)^{1-n}|\hat{\boldsymbol{\Sigma}}|^{-1}\right)\right). \tag{4.30}$$

Recall that the spectrum of a principle submatrix $\hat{\boldsymbol{\Sigma}}$ is bounded by the spectrum of the original matrix Σ. By [41], we may conclude that:

$$|\hat{\boldsymbol{\Sigma}}| = O(\Pi_{i=1}^{n} p_i). \tag{4.31}$$

From equations (4.16) and (4.17) we have that:

$$\frac{1}{\sqrt{\beta}} = \frac{1}{1 - \frac{c}{\sqrt{m}}} \Rightarrow m = \left(\frac{c}{1-\sqrt{\beta}}\right)^2.$$

As (4.20) only holds if $m \geq m_\epsilon$, if we choose c, m in the following manner:

$$\begin{aligned} c &\geq O\left(1 \wedge \log^{\frac{1}{2}}\left(\epsilon^{-2}(2\pi)^{1-n}|\hat{\boldsymbol{\Sigma}}|^{-1}\right) \wedge \sqrt{m_\epsilon}|1-\sqrt{\beta}|\right), \\ &= O\left(1 \wedge \log^{\frac{1}{2}}\left(\epsilon^{-2}(2\pi)^{1-n}\left(\Pi_{i=1}^{n} p_i\right)^{-1}\right) \wedge \sqrt{m_\epsilon}|1-\sqrt{\beta}|\right), \qquad (4.32) \\ m &= \left(\frac{c}{1-\sqrt{\beta}}\right)^2, \qquad (4.33) \end{aligned}$$

we have that $\Pr\left[\boldsymbol{p^r} \geq \frac{1}{\sqrt{\beta}}\hat{\boldsymbol{p}}^{\boldsymbol{r}}\right] \geq (1-\epsilon)$. Repeating the analysis for the column distribution $\boldsymbol{p^c}$, we note that we need to choose the minimum of the two products $\Pi_{i=1}^{n_1} p_i^r, \Pi_{j=1}^{n_2} p_j^c$ in order to have the (4.1) hold. Thus choosing c and m as stated in Lemma 4.2.1, the proof is complete. □

4.3.2 Proof of Lemma 4.2.3

Proof. As in the proof of Lemma 4.2.1, we proceed in the case that $\boldsymbol{p} = \boldsymbol{p^r}, \hat{\boldsymbol{p}} = \hat{\boldsymbol{p}}^{\boldsymbol{r}}$; the analysis for the column case is the same.

By convergence in distribution we have that:

$$\Pr[c_1\boldsymbol{p} \leq \sqrt{m}(\boldsymbol{p} - \hat{\boldsymbol{p}}) \leq c_2\boldsymbol{p}] \xrightarrow{\mathcal{D}} \Pr[c_1\boldsymbol{p} \leq Y \leq c_2\boldsymbol{p}].$$

Therefore, there exists an index $m_\epsilon \in \mathbb{N}$ such that for all $m \geq m_\epsilon$ we have that:

$$\Pr[c_1\boldsymbol{p} \leq \sqrt{m}(\boldsymbol{p} - \hat{\boldsymbol{p}}) \leq c_2\boldsymbol{p}] \geq \Pr[c_1\boldsymbol{p} \leq \boldsymbol{Y} \leq c_2\boldsymbol{p}] - \epsilon/2.$$

Building upon our previous analysis for Lemma 4.2.1, we see that we wish to bound the following integral:

$$I := \frac{1}{\sqrt{(2\pi)^{n-1}|\hat{\boldsymbol{\Sigma}}|}} \int_{c_1 p_1}^{c_2 p_1} \cdots \int_{c_1 p_{n-1}}^{c_2 p_{n-1}} \exp\left(-\frac{1}{2}\left(\frac{1}{p_n}\left(x_1 + \cdots + x_{n-1}\right)^2 + \sum_{i=1}^{n-1} \frac{x_i^2}{p_i}\right)\right) d\boldsymbol{x}_{n-1}, \tag{4.34}$$

below by $1 - \frac{\epsilon}{2}$ with $c_1 < 0$ and $c_2 > 0$. Observe that if we can find constants $c_1 < 0, c_2 > 0$ such that $I \geq 1 - \epsilon/2$, then for all $m \geq m_\epsilon$ we have that:

$$\Pr[c_1\boldsymbol{p} \leq \sqrt{m}(\boldsymbol{p} - \hat{\boldsymbol{p}}) \leq c_2\boldsymbol{p}] \geq (1 - \epsilon). \tag{4.35}$$

Note that by symmetry it suffices to find positive constants $c_1, c_2 > 0$ such that the following two inequalities hold:

$$I_1 := \frac{1}{\sqrt{(2\pi)^{n-1}|\hat{\boldsymbol{\Sigma}}|}} \int_{c_1 p_1}^{\infty} \cdots \int_{c_1 p_{n-1}}^{\infty} \exp\left(-\frac{1}{2}\left(\frac{1}{p_n}\left(x_1 + \cdots + x_{n-1}\right)^2 + \sum_{i=1}^{n-1} \frac{x_i^2}{p_i}\right)\right) dV_{n-1}$$
$$\leq \frac{\epsilon}{4} \tag{4.36}$$

$$I_2 := \frac{1}{\sqrt{(2\pi)^{n-1}|\hat{\boldsymbol{\Sigma}}|}} \int_{c_2 p_1}^{\infty} \cdots \int_{c_2 p_{n-1}}^{\infty} \exp\left(-\frac{1}{2}\left(\frac{1}{p_n}\left(x_1 + \cdots + x_{n-1}\right)^2 + \sum_{i=1}^{n-1} \frac{x_i^2}{p_i}\right)\right) dV_{n-1}$$
$$\leq \frac{\epsilon}{4}. \tag{4.37}$$

From previous analysis, we conclude that to satisfy (4.36) and (4.37), it suffices to choose c_1, c_2 as follows:

$$c_1, c_2 \geq O\left(1 \wedge \log^{\frac{1}{2}}\left(\epsilon^{-2}(2\pi)^{1-n}\left(\Pi_{i=1}^{n} p_i\right)^{-1}\right)\right) \tag{4.38}$$

Observe that the LHS of (4.35):

$$\Pr[c_1 \boldsymbol{p} \leq \sqrt{m}(\boldsymbol{p} - \hat{\boldsymbol{p}}) \leq c_2 \boldsymbol{p}] = \Pr[c_1/\sqrt{m}\boldsymbol{p} + \hat{\boldsymbol{p}} \leq \boldsymbol{p} \leq c_2/\sqrt{m}\boldsymbol{p} + \hat{\boldsymbol{p}}], \tag{4.39}$$

which implies that the following two inequalities hold:

$$\boldsymbol{p} \leq c_2/\sqrt{m}\boldsymbol{p} + \hat{\boldsymbol{p}} \Leftrightarrow \boldsymbol{p}(1 - c_2/\sqrt{m}) \leq \hat{\boldsymbol{p}}$$
$$\hat{\boldsymbol{p}} + c_1/\sqrt{m}\boldsymbol{p} \leq \boldsymbol{p} \Leftrightarrow \hat{\boldsymbol{p}} \leq (1 - c_1/\sqrt{m})\boldsymbol{p}.$$

Combining the above two inequalities yields:

$$\frac{1}{1 - \frac{c_1}{\sqrt{m}}}\hat{\boldsymbol{p}} \leq \boldsymbol{p} \leq \frac{1}{1 - \frac{c_2}{\sqrt{m}}}\hat{\boldsymbol{p}} \tag{4.40}$$

Setting the following terms equal:

$$\frac{1}{\sqrt{\beta_1}} = \frac{1}{1 - \frac{c_1}{\sqrt{m}}} \Rightarrow m = \left(\frac{c_1}{1 - \sqrt{\beta_1}}\right)^2,$$
$$\frac{1}{\sqrt{\beta_2}} = \frac{1}{1 - \frac{c_2}{\sqrt{m}}} \Rightarrow m = \left(\frac{c_2}{1 - \sqrt{\beta_2}}\right)^2.$$

Setting the corresponding terms equal, we see that we must require that:

$$c_1 = -\left|\frac{1 - \sqrt{\beta_1}}{1 - \sqrt{\beta_2}}\right| c_2. \tag{4.41}$$

We choose c_1, c_2, m in the following way:

$$c_1 = -\left|\frac{1-\sqrt{\beta_1}}{1-\sqrt{\beta_2}}\right| c_2, \tag{4.42}$$

$$c_2 = O\left(1 \wedge \log^{\frac{1}{2}}\left(\epsilon^{-2}(2\pi)^{1-n}\left(\min\left(\Pi_{i=1}^{n_1} p_i^r, \Pi_{j=1}^{n_2} p_j^c\right)\right)^{-1} \wedge \sqrt{m\epsilon}|1-\sqrt{\beta_2}|\right)\right), \tag{4.43}$$

$$m = \left(\frac{c_2}{1-\sqrt{\beta_2}}\right)^2. \tag{4.44}$$

Note that if we choose c_1, c_2 according to (4.42) and (4.43) then (4.36) and (4.37) are satisfied. Choosing m according to (4.44), then (4.35) is satisfied. If (4.35) holds, then with probability at least $(1-\epsilon)$ we have that:

$$\frac{1}{1-\frac{c_1}{\sqrt{m}}}\hat{\boldsymbol{p}} \leq \boldsymbol{p} \leq \frac{1}{1-\frac{c_2}{\sqrt{m}}}\hat{\boldsymbol{p}} \Leftrightarrow \frac{1}{\sqrt{\beta_1}}\hat{\boldsymbol{p}} \leq \boldsymbol{p} \leq \frac{1}{\sqrt{\beta_2}}\hat{\boldsymbol{p}}$$

Repeating the analysis with the column distribution, we conclude that (4.4) holds. □

4.4 Matrix Completion Guarantees

With Lemma 4.2.1 in hand, we are prepared to prove Theorem 4.2.2 in Section 4.4.1. In Section 4.4.2, we will extend the reasoning in [3] which quantifies the relaxation of the conditions for which (3.10) succeeds in obtaining exact recovery using the empirically learned weights.

4.4.1 Proof of Theorem 4.2.2

Proof. Assume that $\boldsymbol{R}, \boldsymbol{C}$ are a set of weights satisfying (4.3). For any arbitrary $\epsilon \in (0,1)$, if we choose m, c as in (4.32) and (4.33), recalling that $\hat{\boldsymbol{p}}$ is

an empirical estimator of $\frac{1}{\sum_{i,j} p_{ij}}\boldsymbol{p}$, we have that for all $(i,j) \in [n] \times [n]$ with probability at least $(1-\epsilon)^2$:

$$\begin{aligned}
\frac{p_{ij}}{\sum_{i,j} p_{ij}} &\geq \frac{1}{\beta}\hat{p}_{ij} \\
\Leftrightarrow p_{ij} &\geq \frac{\sum_{i,j} p_{ij}}{\beta}\hat{p}_{ij} \\
&\geq c_0 \left(\frac{R_i^2}{\sum_{i' \in \mathcal{S}_r} R_{i'}^2} + \frac{C_j^2}{\sum_{j' \in \mathcal{S}_c} C_{j'}^2} \right) \log^2(2n). \qquad (4.45)
\end{aligned}$$

We observe that (4.45) is equivalent to (3.11) without the monotonicity assumption. Therefore Theorem 3.3.2 applies. As the stage one sampling and stage two sampling are independent, (3.10) attains exact recovery with probability at least $(1 - 5(2n)^{-10})(1-\epsilon)^2$. Thus the proof is complete. □

4.4.2 Weighted Nuclear Norm and Relaxation of Sufficient Recovery Conditions

With Theorem 4.2.2 we established some sufficient conditions for the weights $\boldsymbol{R}, \boldsymbol{C}$ in order for (3.10) to attain exact recovery. In this section we will establish exact recovery guarantees for a specific set of weights defined in terms of the empirical sampling distribution $\hat{\boldsymbol{p}}$ and quantify how the exact recovery conditions for (3.10) are relaxed relative to unweighted nuclear norm minimization (3.4). The weights we will be concerned with will be the empirical analogue of weights defined in [3].

4.4.2.1 Proof of Theorem 4.2.4

We now prove that a specific choice of weights which are the direct empirical analogue of weights posed in [3] produces a quantifiable relaxation in exact recovery conditions.

Proof. Choosing the weights $\boldsymbol{R}, \boldsymbol{C}$ as in (4.8) and (4.9), observe that for any $(i,j) \in [n] \times [n]$:

$$\left(\frac{R_i^2}{\sum_{i' \in \mathcal{S}_r} R_{i'}^2} + \frac{C_j^2}{\sum_{j' \in \mathcal{S}_c} C_{j'}^2} \right) \log^2(2n) = \left(\frac{\hat{p}_i^r \sum_{j' \in \mathcal{S}_c} \hat{p}_{j'}^c + \hat{p}_j^c \sum_{i' \in \mathcal{S}_r} \hat{p}_{i'}^r}{\sum_{i',j' \in \mathcal{S}_r, \mathcal{S}_c} \hat{p}_{i'}^r \hat{p}_{j'}^c} \right) \log^2(2n). \tag{4.46}$$

For arbitrary $\epsilon \in (0,1)$ and for some β_1, β_2, suppose that we have that for all indices $(i,j) \in [n] \times [n]$, with probability at least $(1-\epsilon)^2$:

$$\frac{1}{\beta_1} \hat{p}_{ij} \leq \frac{p_{ij}}{\sum_{i,j} p_{ij}} \leq \frac{1}{\beta_2} \hat{p}_{ij}. \tag{4.47}$$

Applying (4.47) to (4.46) we have that for any $(i,j) \in [n] \times [n]$:

$$\begin{aligned}
\left(\frac{R_i^2}{\sum_{i' \in \mathcal{S}_r} R_{i'}^2} + \frac{C_j^2}{\sum_{j' \in \mathcal{S}_c} C_{j'}^2} \right) \log^2(2n) &= \left(\frac{\hat{p}_i^r \sum_{j' \in \mathcal{S}_c} \hat{p}_{j'}^c + \hat{p}_j^c \sum_{i' \in \mathcal{S}_r} \hat{p}_{i'}^r}{\sum_{i',j' \in \mathcal{S}_r, \mathcal{S}_c} \hat{p}_{i'}^r \hat{p}_{j'}^c} \right) \log^2(2n) \\
&\leq \frac{\beta_1}{\beta_2} \left(\frac{p_i^r \sum_{j' \in \mathcal{S}_c} p_{j'}^c + p_j^c \sum_{i' \in \mathcal{S}_r} p_{i'}^r}{\sum_{i',j' \in \mathcal{S}_r, \mathcal{S}_c} p_{i'}^r p_{j'}^c} \right) \log^2(2n) \\
&= \frac{\beta_1}{\beta_2} \left[\frac{p_i^r \log^2(2n)}{\sum_{i' \in \mathcal{S}_r} p_{i'}^r} + \frac{p_j^c \log^2(2n)}{\sum_{j' \in \mathcal{S}_c} p_{j'}^c} \right] \\
&\leq \frac{\beta_1}{\beta_2} \left[\frac{p_i^r \log^2(2n)}{\sum_{i' \in \mathcal{S}_r^*} p_{i'}^r} + \frac{p_j^c \log^2(2n)}{\sum_{j' \in \mathcal{S}_c^*} p_{j'}^c} \right] \\
&\leq \frac{\beta_1}{\beta_2} \frac{1}{c_0} \mathfrak{C} 2 p_{ij}, \qquad (4.48) \\
&\leq \frac{\beta_1}{\beta_2} \frac{1}{2c_0} p_{ij}. \qquad (4.49)
\end{aligned}$$

where (4.48) follows from (4.10) and (4.11). Now we may choose β_1, β_2 such that $1 \leq \frac{\beta_1}{\beta_2} \leq 2$, then (4.49) immediately yields that for any $(i,j) \in [n] \times [n]$:

$$p_{ij} \geq c_0 \left(\frac{R_i^2}{\sum_{i' \in \mathcal{S}_r} R_{i'}^2} + \frac{C_j^2}{\sum_{j' \in \mathcal{S}_c} C_{j'}^2} \right) \log^2(2n), \tag{4.50}$$

which is exactly equivalent to (3.11) without the monotonicity assumption. Thus to complete the proof, we must ensure that (4.47) holds. We may apply Lemma 4.2.3 to conclude that for β_1, β_2 chosen as above, and $\mathcal{C}$ as in (4.10) and (4.11), if we choose m, c_1, c_2 as in (4.5)-(4.7), we have that (4.47) holds with probability at least $(1-\epsilon)^2$ and the proof is complete. $\square$

Chapter 5

Conclusion

In this thesis we explored how to efficiently incorporate two different types of weights, one class of weights coming from expert or prior knowledge and another class of weights which we learned from some samples in the absence of any sort of prior knowledge, into optimization techniques. In both cases we presented theoretical guarantees of these weighted optimization techniques and we presented numerical experiments which provide evidence for the claim that the appropriate use of weights can allow for a simultaneous reduction in sample complexity and an improvement in approximation accuracy.

Bibliography

[1] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser Mathematics, 2013.

[2] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703, November 2009.

[3] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward. Completing Any Low-rank Matrix, Provably. *ArXiv e-prints arXiv:1306.2979v4*, 2014.

[4] D.L. Ruderman and W. Bialek. Statistics of natural images: Scaling in the woods. *Phys. Rev. Lett.*, 73(6), 1994.

[5] D.J. Tolhurst, Y. Tadmor, and T. Chao. Amplitude spectra of natural images. *Ophthalmic and Physiological Optics*, 12(2):229–232, April 1992.

[6] N. Silver. *The Signal and the Noise: Why So Many Predictions Fail but Some Don't*. Penguin Press.

[7] H. Rauhut and R. Ward. Interpolation via weighted ℓ_1 minimization. *arXiv:1308.0759*, August 2013.

[8] M. Khajehnejad, W. Xu, A. Avestimehr, and B. Hassibi. Analyzing

weighted ℓ_1 minimization for sparse recovery with nonuniform sparse models. *IEEE Transactions on Signal Processing*, 59(5):1985–2001, 2011.

[9] A. Krishnaswamy, S. Oymak, and B. Hassibi. A simpler approach to weighted ℓ_1 minimization. *ICASSP 2012*, pages 3621–3624.

[10] S. Misra and P. Parrilo. Analysis of weighted ℓ_1-minimization for model based compressed sensing. *arXiv:1301.1327 [cs.IT]*, 2013.

[11] S. Schwartz and A. Tewari. Stochastic methods for ℓ_1 regularized loss minimization. *Journal of Machine Learning Research*, 12:1865–1892, June 2011.

[12] T. Blumensath and M. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, November 2009.

[13] R. L. Plackett. Some theorems in least squares. *Biometrika*, 37(1/2):pp. 149–157, 1950.

[14] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, April 1995.

[15] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[16] J. Tropp and A. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655 – 4666, 2007.

[17] D. Needell and J. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, May 2009.

[18] Mathematica version 8.0, 2010.

[19] T. Blumensath and M. Davies. Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14(5-6):629–654, December 2008.

[20] K. Lange. *Optimization*. Springer Verlag, 2004.

[21] T. Blumensath and M. Davies. Normalized iterative hard thresholding: Guaranteed stability and performance. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):298–309, April 2010.

[22] Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. on Optimization*, 20(4):1956–1982, March 2010.

[23] Emmanuel Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Commun. ACM*, 55(6):111–119, June 2012.

[24] Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theor.*, 56(5):2053–2080, May 2010.

[25] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inf. Theor.*, 57(3):1548–1566, March 2011.

[26] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 665–674, 2013.

[27] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998, 2010.

[28] Sahand Negahban and Martin J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13.

[29] Benjamin Recht. A simpler approach to matrix completion. *J. Mach. Learn. Res.*, 12:3413–3430, December 2011.

[30] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, August 2010.

[31] Ruslan Salakhutdinov and Nathan Srebro. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. arxiv.org/abs/1002.2780, 2010.

[32] Rina Foygel, Ruslan Salakhutdinov, Ohad Shamir, and Nati Srebro. Learning with the weighted trace-norm under arbitrary sampling distributions. *NIPS Proceedings*, 24, 2011.

[33] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[34] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[35] Jon A. Wellner A. W. van der Vaart. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics, 1996.

[36] C.G. Khatri M. S. Srivastava. *An Introduction to Multivariate Statistics*. Elsevier North Holland Inc., 1979.

[37] K. R. Parthasarathy. *Probability Measures on Metric Spaces*. Academic Press Inc., 1967.

[38] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[39] Hanna K. Jankowski and Jon A. Wellner. Estimation of a discrete monotone distribution. *Electronic Journal of Statistics*, 3:1567–1605, 2009.

[40] Shayle R. Searle. *Linear Models*. Wiley-Interscience, 3 1997.

[41] G. S. Watson. Spectral decomposition of the covariance matrix of a multinomial. *Journal of the Royal Statistical Society: Series B*, 58(1):289–291, 1996.

[42] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal On Scientific Computing*, 20:33–61, 1998.